Developments in the Built Environment 14 (2023) 100167

Contents lists available at ScienceDirect

Developments in the Built Environment

journal homepage: www.sciencedirect.com/journal/developments-in-the-built-environment

Quantifying the reliability of defects located by bridge inspectors through human observation behavioral analysis

Pengkun Liu, Ying Shi, Ruoxin Xiong, Pingbo Tang *

Department of Civil and Environmental Engineering, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, 15213, United States

ARTICLE INFO

Keywords:
Bridge inspection
Visual inspection
Defects localization
Process mining
Crowdsourcing
Ensemble learning

ABSTRACT

Distinct site conditions and individual expertise contribute to the subjective nature of bridge inspection processes, which involve uncertain human factors. Assessing inspection reliability can be achieved by examining inspectors' behaviors that lead to inaccurately identified or overlooked defects. However, the scarcity of comprehensive behavioral data regarding defect observation poses a challenge in evaluating inspection consistency. This paper investigates observation behaviors in bridge inspection to quantify the reliability of structural defect localizations. We employ defect inspection strategies correlated with more dependable defect localization records to construct a behavioral process graph that quantifies inspectors' performance and predicts their "inspection reliability index." The generated reliability index for inspectors serves as a weighting factor to emphasize the opinions of more reliable inspectors when consolidating inspection records. The findings reveal that aggregating the inspection records of 96 human subjects based on their reliability indices effectively filters out false alarms while retaining reliable defect records.

1. Introduction

Bridges are essential transportation infrastructure, facilitating daily life and socio-economic activities. To maintain the functionality of transportation networks and public safety, bridge management agencies prioritize bridges for inspection and maintenance (I&M) based on damage severity, budget, and resource constraints. Accurate bridge condition ratings are critical for allocating timely I&M to urgent cases and conserving resources on less urgent cases. Unreliable ratings may mislead I&M plans, leading to resource waste and leaving some bridges in hazardous states. According to a Federal Highway Administration investigation, 95% of primary condition ratings for bridge elements vary within two rating points of the average, and 68% vary within one rating point (Phares et al., 2004).

Condition ratings fluctuate due to the combined effects of human factors, environmental conditions, and the complexity of structures and inspection tasks. This process entails human visual inspection behaviors interacting with environmental conditions and structural complexity (Phares et al., 2004). Engineers may overlook defects and produce unreliable conclusions when unfavorable field conditions and data volume is significant. For instance, between 2009 and 2019, most bridge failures in China were related to human factors (69.6%), far exceeding natural factors (30.4%) (Tan et al., 2020). In 2021, a 40-year unnoticed defect in gusset plates bowing under stress caused the Mississippi River Bridge Collapse (Salem and Helmy, 2014). An engineer photographed this defect as early as 2003, but the decision-makers diagnosis process underestimated its significance. In such cases, inspectors' underlying behaviors using field data appear to have unaccounted issues affecting the reliability of inspection results. Consequently, bridge inspection processes and defect localization are unreliable and have limitations in guiding effective maintenance planning.

Reliable defect detection at the element and structure levels is the basis for reliable ratings. Localization is specifically mentioned in the context of bridge inspection performance because it is a critical aspect that has significant implications for the inspection process's accuracy, safety, and efficiency. Many choices made by inspectors in the defect detection processes can influence the data quality and the derived condition ratings. Any inspection process can be unique in human, sensing, and computing details; controlling all those factors for quantifying how they influence the reported defect reports' reliability is challenging. Without that knowledge, inspectors could not proactively adjust their behaviors to improve defect detection accuracy.

* Corresponding author.
E-mail addresses: pengkunl@andrew.cmu.edu (P. Liu), yingshi@andrew.cmu.edu (Y. Shi), ruoxinx@andrew.cmu.edu (R. Xiong), ptang@andrew.cmu.edu (P. Tang).

https://doi.org/10.1016/j.dibe.2023.100167
Received 7 January 2023; Received in revised form 8 April 2023; Accepted 27 April 2023
Available online 5 May 2023

Understanding the influence of human inspection behaviors on defect localization performance can potentially achieve more reliable bridge inspections with accurate crack locations. The absence of explicit guidelines for visual inspection tactics forces inspectors to rely on subjective judgments or assessment criteria, leading to inconsistent or less reliable defect location assessment results (Juran and De Feo, 2010; Laofor and Peansupap, 2012). Previous studies indicate that experienced inspectors tend to produce more consistent inspection results (Carver, 2003; Ball et al., 2017; Megaw and Richardson, 1979; Woodcock, 2014). Existing research demonstrates the potential of eye-tracking technology or Building Information Modeling (BIM) event log process mining techniques to understand engineers' behaviors (Xu et al., 2019; Wang et al., 2018). Some behavioral studies in other domains suggest that behavioral analysis can estimate inspector reliability and emphasize the results of more reliable inspectors to improve the reliability of aggregated inspection results (Carver, 2003; Ball et al., 2017; Megaw and Richardson, 1979; Woodcock, 2014). Defect inspection strategy in bridge inspection refers to the approach or method employed by inspectors to identify, assess, and prioritize defects in bridge elements during the inspection process. These strategies are crucial for detecting and addressing the most critical defects to maintain structural integrity and safety. Defect inspection strategies may differ among inspectors depending on their experience, knowledge, and understanding of bridge structures. Unfortunately, these studies have not yet explored the relationship between inspection performance and defect inspection strategy for explaining unreliable inspection records and revealing strategies for improving their reliability.

This paper examines human inspection of behavioral analysis for quantifying the reliability of structural defect localizations to test the hypothesis that "aggregating the defect localization results of bridge inspectors with weighted votes can achieve more reliable defect localization." The research questions are:

1) How to capture detailed inspection process behaviors of inspectors?
2) What defect inspection strategy of inspectors can help predict reliable defect localization inspection outcomes?
3) How to quantify the reliability of element-level defects localized by inspectors based on inspection process behaviors?

Some technical challenges form barriers to answering these three research questions. First, capturing detailed inspection process behaviors of inspectors in various contexts of structural condition assessment is difficult because field scenarios could hardly allow the installation of sensors on human bodies to track their behaviors. Second, no process data analytics method exists for extracting meaningful inspection strategies from behavioral process data (e.g., eye-tracking results that trace the attention-transferring processes of bridge inspectors). Third, little effort has been made to establish a solid statistical approach for quantifying the reliability of inspection records based on the historical performance of inspectors.

The proposed research uses bridge inspection gamification as the basis for developing a computational framework that captures and utilizes inspection process behaviors in quantifying the reliability of bridge inspection records. The authors designed a digital bridge inspection game integrating information from the FEM model (stress and displacements with defect simulation) and inspection reports (bridge basic information) to accurately represent and collect inspectors' behaviors and inspection results (defect inspection). The paper then investigates defect inspection strategies associated with more reliable defect localization records. Subsequently, the paper constructs a behavioral process graph based on these effective defect inspection strategies. A behavioral process graph visually represents the sequence of actions, decisions, and interactions that occur during a specific process. It illustrates the connections and relationships between different steps in the process, enabling a better understanding of the behaviors and strategies employed by inspectors. The graph can help identify patterns and trends that associated with more reliable or less reliable outcomes, such as which defect inspection strategies lead to more accurate defect localization records. Lastly, the paper introduces the concept of an "inspection reliability index." This index quantifies inspectors' performance by predicting the likelihood of reliable inspection outcomes based on their observed behaviors and their inspection strategies. Inspectors with higher inspection reliability indices are expected to produce more accurate and reliable defect localization records. In summary, through process mining, crowdsourcing, and ensemble learning, the authors discover inspection strategies from the inspectors' behavioral data to calibrate the bridge element-level defect localization results through inspection behavior analysis.

Three specific research objectives related to these questions include: (1) designing a digital bridge inspection game for accurately representing and collecting inspectors' behaviors and inspection results (defect localization inspection); (2) discovering defect inspection strategies from the inspectors' behavioral data; and (3) quantifying the reliability of element-level defects localized by multiple game participants with diverse inspection capabilities through inspection behavior analysis. The first objective addresses the challenge of capturing detailed inspection process behaviors. The second objective aims to overcome the challenge of extracting meaningful inspection strategies from behavioral process data. The third objective resolves the challenge of quantifying the reliability of defects' location records generated by multiple inspection game participants.

The following sections are organized: Section 2 presents a motivating case for the reliability problem of bridge inspection. Section 3 demonstrates the results of the literature review. Section 4 proposes the framework and methodology of bridge inspection gamification, collection, and process mining for the inspectors' behaviors. A validation study in Section 5 discusses the results of quantifying the reliability of bridge defect localization using inspection behavior analysis. Sections 6 and 7 conclude the study with significant findings and remarks about future research directions.

2. Motivation case

The purpose of the motivating case is to provide a visual and intuitive explanation of the inspection behavior differences and how those behaviors could be useful for quantifying the reliability of defect localization of the human. Such an illustration needs a visualization without getting into details of various interpretations of the observed actions and contextual events.

A bridge inspection game, informed by a literature review, enabled the authors to identify challenges and opportunities associated with predicting the reliability of bridge defect localization based on inspection behavior analysis. This game was a motivating case to illustrate the need to analyze inspection process behaviors to better understand the uncertainties in bridge inspection processes and defect records (Liu et al., 2021). The bridge inspection game allows "players" (bridge inspectors) to assess bridge defects and underlying causes using finite element model (FEM) simulation data (stress and displacements) and inspection reports (basic bridge information and potential defect types) (Liu et al., 2021). The bridge inspection reports and data are derived from a continuous rigid frame bridge (CRFB). During the game, the "players" examine each bridge element for various defects within a 15-min timeframe.

The authors designed two FEMs representing two possible conditions of the studied CRFB to depict "good" and "defective" conditions, as shown in Figs. 1 and 2. The original CRFB model represents the "good condition" without any stiffness reductions (Fig. 1). To simulate damage, the alternative CRFB model features stiffness reduction in the box-girders at the mid-span (Fig. 2). Stress distributions differed between the two models due to stiffness reductions that simulated bridge defects, particularly visible in Fig. 2 (b), (c), and Fig. 3 (b), (c). The eye-tracking device captured the "players'" behaviors when identifying bridge

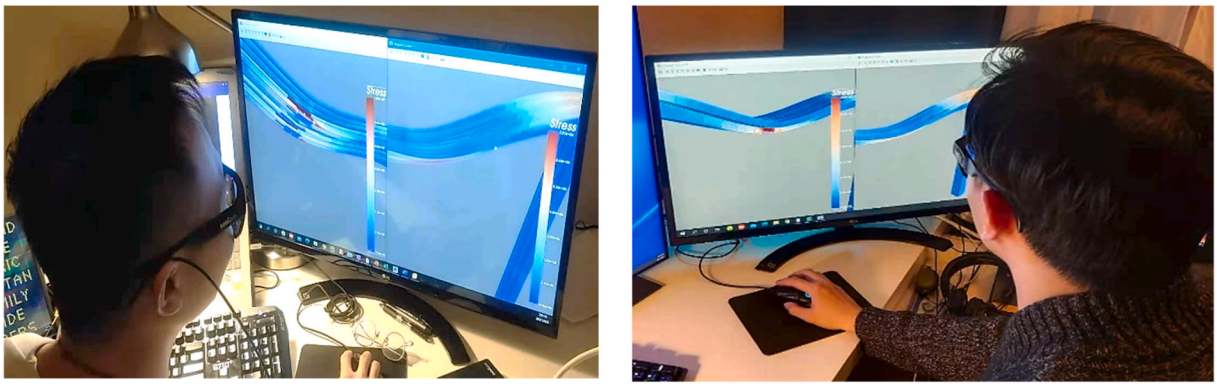


Fig. 1. Bridge inspection behavior collections with eye-tracking.

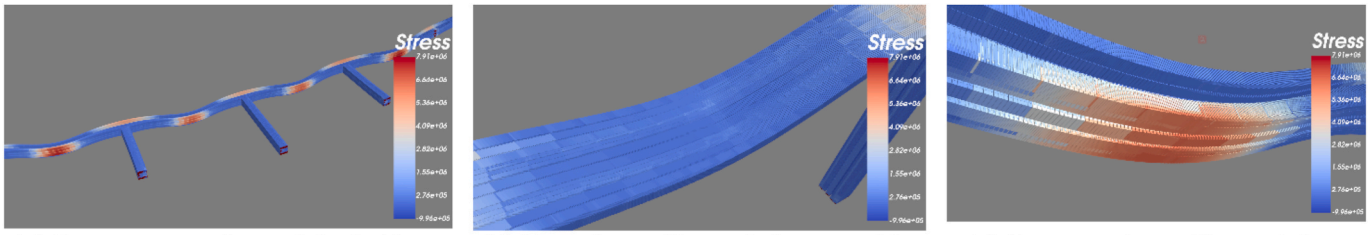


Fig. 2. FEM of CRFB without stiffness reduction represents "good" conditions.

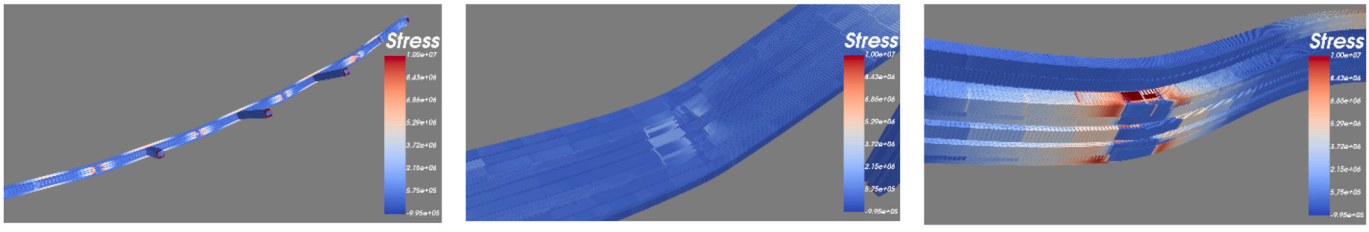


Fig. 3. FEM of CRFB with stiffness reductions on box-girders of mid-span represent "defective" conditions (Liu et al., 2021).

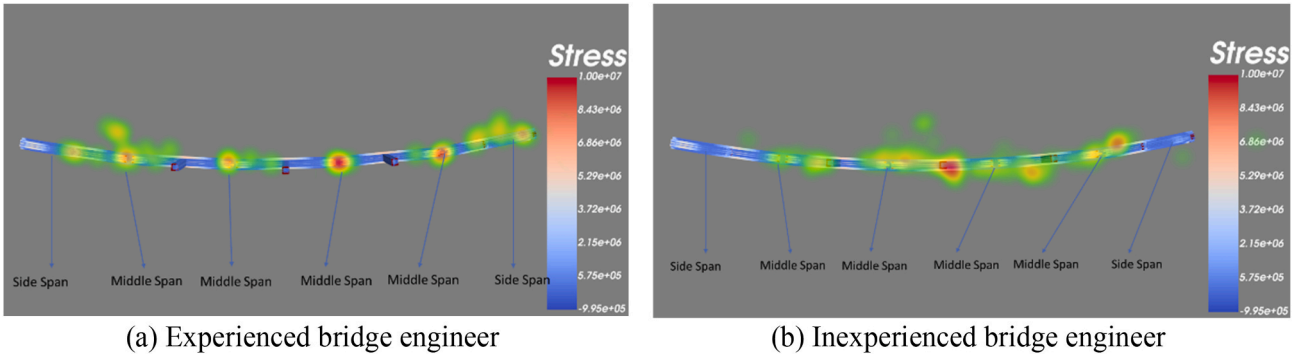


Fig. 4. Heat map of the inspectors' gazes for finding the defects on bottom slabs.

defects, including eye movements and sequences of selections, as shown in Fig. 1.

The eye movements of inspectors can reveal their attention-shifting processes, which may be used to infer their inspection skill levels. As illustrated in Fig. 4, engineers' gaze movements can help deduce their search strategies during the inspection. Two groups of human subjects participated in this game to determine whether structural engineering knowledge and experience influence their inspection behaviors. Specifically, when detecting defects on the bridge's underside, graduate students with structural engineering backgrounds focused more on the

bottom slab at each mid-span. They tended to search for similar locations and defects between spans, as shown in Fig. 4 (a). Consequently, the gaze movement trajectory exhibits a horizontal elongation pattern along the bridge axis. In contrast, participants with limited structural engineering knowledge employed a random search approach. As depicted in Fig. 4 (b), these individuals' average gaze movement trajectory is more clustered and spherical than experienced engineers.

These initial findings published in (Liu et al., 2021) inspired the authors to extract valuable defect-searching strategies from inspection process data. We found that eye-tracking and mouse-click data serve different purposes and provide different insights when visualizing or studying user behavior. While mouse-click data can be more convenient and cost-effective to collect, eye-tracking data serve a better visualization purpose. Eye-tracking technology can reveal the attention-shifting processes of various inspectors, helping to infer and understand their defect inspection search strategies. However, during the collection of eye-tracking data, the authors observed that although eye-tracking data offers several advantages, such as providing more detailed information about user behavior (where users are looking, how long they spend looking at specific areas, and the order in which they view different bridge elements on a screen), the use of eye-tracking equipment is labor-intensive and impractical for collecting much human behavior data. Therefore, the authors designed a bridge inspection game for accurately representing and collecting the inspectors' behaviors and inspection results. Because we could distribute this digital game and collect mouse click data at the same time on several devices.

3. Literature review

To address the research question, this section reviews several categories of relevant studies: 1) inspection behavior capturing and uncertainty analysis, 2) extracting and assessing inspection strategies from inspection behavior process data, and 3) quantifying the reliability of defects identified by inspectors.

3.1. Inspection behavioral capturing and uncertainty analysis

This subsection examines studies on capturing behavioral processes to collect data and understand human behaviors. Inspection behavioral capturing is an emerging research area focusing on recording and analyzing human inspectors' behavior during various inspection tasks. By understanding human inspectors' decision-making processes and strategies, researchers can develop more efficient, reliable, and effective inspection systems.

3.1.1. Uncertainty analysis in inspection

Uncertainty analysis is essential for evaluating the reliability of inspection results and supporting informed decision-making. Uncertainty in inspection processes arises from various sources, including measurement errors, environmental conditions, human factors, and the inherent variability of inspected objects (Iso and OIML, 1995). Understanding and identifying the sources of uncertainty in inspection processes is critical for developing strategies to mitigate their impact on inspection outcomes. Human factors play a significant role in introducing uncertainties in inspection processes. Studies by Drury et al. (Drew et al., 2013) explored the influence of human factors on inspection performance and uncertainties, highlighting the need for a comprehensive understanding of human factors to improve inspection reliability. Currently, bridge inspection predominantly relies on visual inspection. The visual inspection entails a thorough and critical assessment of an object, referring to a predefined standard. During visual inspection, inspectors must mentally process, focus on, and relay information, utilizing both short-term and long-term memory (Gallwey and Drury, 1986). The absence of explicit guidelines for visual inspection tactics forces inspectors to rely on their subjective judgments or assessment criteria, leading to inconsistent or less reliable defect location assessment results (Juran and De Feo, 2010; Laofor and Peansupap, 2012).

3.1.2. Behavioral capturing in inspection

Understanding and analyzing inspectors' behavior during inspection can provide valuable insights for improving inspection performance and reducing errors. Various studies have focused on different aspects of behavioral inspection capturing:

Various methods have been proposed to capture inspection behavior, providing insights into human inspectors' strategies and decision-making processes. Current research demonstrates the potential for using eye-tracking or BIM event log process mining techniques to understand engineers' behaviors (Xu et al., 2019; Wang et al., 2018). Eye-tracking has been extensively used in inspection studies to understand visual attention and search patterns during inspections. Human eye movements (fixation count, total fixation time, and fixation duration captured through eye-tracking technology) or command records (Pan and Zhang, 2020a, 2020b, 2021a, 2021b; Pan et al., 2020) can help infer individuals' detailed observations and cognitive processes. Mouse-tracking is another approach to capturing inspection behavior, especially in computer-based inspections. Studies by Huang et al. (2012) employed mouse-tracking to understand decision-making and user interaction patterns during software inspections and web-based inspections, respectively. Moreover, in other domains like software development, research has shown that inspectors' performance varies significantly, even when using the same inspection technique; this variation often results from inherent differences among the inspectors employing the technique (Carver, 2003). The aforementioned studies have not explored the relationship between human performance and behavioral processes to explain unreliable records and uncover strategies to enhance reliability.

3.2. Extracting and assessing inspection strategies from inspection behavioral process data

This subsection reviews studies on the challenge of lacking a process data analytics method for extracting meaningful inspection strategies from behavioral process data. Many domains use "Process Mining" to discover strategies of experienced workers, such as design strategies and strategies for inspecting and controlling engineering systems. Those studies show the potential of discovering defect-searching strategies of bridge inspectors.

Process mining aims to generate the process model accurately describing the as-happened processes from the event logs (Buijs et al., 2012). Currently, there are several process discovery algorithms (Gomes et al., 2021), including the alpha miner algorithm (Van der Aalst et al., 2004), heuristic miner algorithm (Weijters et al., 2006), inductive miner algorithm (van der Aalst, 2010), and fuzzy miner algorithm (van der Aalst, 2010). The alpha miner algorithm is the first process discovery algorithm aiming at reconstructing causality from a set of sequences of events following a certain order and showing the results in the Petri net diagram (Gomes et al., 2021). It converts the event logs into direct follows, sequence (causality), parallel and choice relations to create a Petri net describing the process model. However, it does not consider frequencies. The heuristic miner algorithm is improved based on the alpha miner algorithm. It could abstract from exceptional behavior, noise, and low structured data, which is more robust with the semantics of splits and joins (Weijters et al., 2006). The inductive miner algorithm relies on detecting various cuts on the directly follows graph created using the event log to handle infrequent behavior well and finishes quickly for infrequent behavior and large event logs (Leemans et al., 2013). The fuzzy miner algorithm is the first to directly address the problems of large numbers of activities and highly unstructured behavior suitable for the complex and unstructured log data (Günther and Van Der Aalst, 2007).

Overall, process mining methods discover, monitor, and suggest

process improvements by extracting knowledge from design, engineering, or operational information systems' event logs (Pan and Zhang, 2020a, 2020b, 2021a, 2021b; Pan et al., 2020; Van Der Aalst, 2012; Wu et al., 2021a, 2021b, 2022, 2023). It is a discipline providing compressive fact-based insights from actual event logs and supports process improvements or automates the process of extracting and mining process-related information (Van der Aalst, 2016). More specifically, valuable insights from key events, diagnosis failure causes, or repetitive patterns could be detected by interpreting the logs through process mining (Pan and Zhang, 2021a). Process mining has taken essential roles in different domains, such as healthcare (Rojas et al., 2016; Mans et al., 2008), business (Van Der Aalst et al., 2007; Tiwari et al., 2008), education (Bogarín et al., 2018; Cairns et al., 2015), and so on.

Recent civil engineering research on mining BIM logs has emerged (Pan and Zhang, 2020a, 2020b, 2021a, 2021b; Pan et al., 2020; Chua and Hossain, 2011; Al Hattab and Hamzeh, 2018; Kouhestani and Nik-Bakht, 2020). Kouhestani and Nik-Bakht studied capturing event logs and analyzed the process for the design authoring phase of building projects to identify measures derived from the designed process, which could guide the manager to monitor, control, and reset the design works (Kouhestani and Nik-Bakht, 2020). Pan and Zhang proposed automated process discovery from event logs to understand the actual progress of the construction project (Pan and Zhang, 2021a). In addition, Pan and Zhang developed a clustering-based BIM event log mining method to discover the knowledge of design productivity characteristics (Pan and Zhang, 2020a). However, all these studies on BIM log mining have been limited to the design and construction stages. Furthermore, few analyses have yet considered the BIM logs to evaluate the behaviors of the participants to understand their performances.

3.3. Quantification of the reliability of inspectors' decisions

This section reviews studies related to the challenge of establishing a rigid statistical approach for quantifying the inspection records' reliability based on inspectors' historical performance. Some behavioral analysis studies analyzed human data collection and analysis behaviors to identify behaviors and processes that lead to more reliable data collection and analyses (Zheng et al., 2020; Liu et al., 2015; Wang et al., 2019). Evaluating interrater reliability (IRR) is a common objective of many human-related research studies (Gisev et al., 2013; Remenyi et al., 2019; Fletcher et al., 2011). Kappa calculations from Cohen's work are one of the original and most commonly used IRR indices (Cohen, 1960). The extended Fleiss' kappa was developed by Fleiss for use when nominal categories are assessed by multiple humans (Fleiss, 1971). It is widely used in the targets being rated (e.g., patients in medical practice, learners taking a driving test, customers in a shopping mall/center, and burgers in a fast-food chain).

"Crowdsourcing" and "Ensemble Learning" methods integrated with human behavioral analysis can automatically predict less reliable parts of the data or data analysis results based on the behavioral analysis results.

Crowdsourcing is an emerging and powerful approach for collecting data and information and making better decisions based on aggregating knowledge at large scales instead of individuals (Zheng et al., 2020; Liu et al., 2015; Wang et al., 2019; Howe, 2006). However, the challenge in crowdsourcing is the quality control of the final results or labels because the crowd workers have different levels of experience or domain knowledge (Liu et al., 2015; Wang et al., 2019). Furthermore, humans make decisions often with significant individual subjective judgments (Liu et al., 2015). The most common way for crowdsourcing is to make labels or annotations from different workers and use a majority vote to infer the defective locations agreed on by most workers. However, the underlying assumption for the majority vote is that all workers have the same abilities and share the same possibility of making errors (Oyama et al., 2013). Every worker has a unique background, experience, and abilities, Hence, treating the labels generated by different inspectors equally is not efficient and reliable for getting defective locations with high reliability due to a mixture of diverse inspectors. Some researchers have examined inferring the truths of label results by considering the differences in abilities between workers, such as using the expectation-maximization algorithm or confidence scores (Liu et al., 2015; Oyama et al., 2013; Drapeau et al., 2016). The possibility of using different confidence scores to improve the quality of crowdsourcing labels was investigated because the reported difficulty correlates with the probability of a correct answer (Kazai, 2011). However, some workers may be overconfident with their judgments and report higher confidence even though their judgments are wrong. In comparison, others may be underconfident in their judgments and report lower confidence even though their judgments are correct. Therefore, it is significant to calibrate the crowdsourcing results by aggregating the ground truth labels or information from the domain experts (Liu et al., 2015).

On the other hand, ensemble learning is a machine learning paradigm where multiple models (often called "weak learners") are trained to solve the same problem and combined to get better results. The hypothesis is that we can obtain more accurate and/or robust models when weak models are correctly combined. In ensemble learning theory, weak learner models can be used as building blocks for designing more complex models by combining several of them. Most of the time, these basic models perform poorly by themselves either because they have a high bias or too much variance to be robust. Then, the idea of ensemble methods is to reduce the bias and variance of such weak learners by combining several of them to create a strong learner (or ensemble model) that achieves better performance.

Several types of research reveal the challenges of reliable data collection with crowdsourcing (Zheng et al., 2020; Liu et al., 2015; Wang et al., 2019). Wang et al. proposed a new crowdsourcing approach for labeling images of complex construction scenes with safety-rule violations by using a Bayesian network-based crowd consensus model aggregating the labels from annotators to obtain reliable safety-rule violation labels (Wang et al., 2019). Liu et al. framed the optimal allocation of true labels to calibrate the crowdsourced labels as the submodular optimization problem with a greedy allocation strategy, which encourages acquiring true labels for the most uncertain items (Liu et al., 2015). Zhong et al. proposed a temporally calibrated method for identifying the spatial-temporal distribution of PM2.5 in intra-urban areas (Zheng et al., 2020). Oyama et al. developed a method for using the confidence scores from the crowdsourcing workers to integrate labels because some workers are confident with their labels while others are underconfident (Oyama et al., 2013). Zhang proposed a novel general ensemble method for learning from crowds, which does not infer the true labels of training instances and directly builds learning models from the crowdsourced labeled data to predict class labels of unlabeled instances (Zhang et al., 2018).

3.4. Research gaps and contributions of the proposed new methods

Unique site conditions and personal knowledge lead to the "subjective" nature of bridge inspection processes containing uncertain human factors. One way to assess inspection reliability is to analyze inspectors' behaviors leading to incorrectly located or missing defects. A review of related studies reveals three research gaps in quantifying the reliability of defect localizations from the visual bridge inspection process. First, bridge inspection largely depends on inspectors' subjective judgments, and evidence is lacking to understand inspection strategies and evaluate inspectors' performance. Second, there is a dearth of process data analytics methods for extracting inspection strategies from behavioral data. Third, the unreliability of element-level defect localizations arises from the fact that they are typically generated and integrated by multiple inspectors with varying experience and abilities, making it challenging to assess their reliability based on inspectors' behaviors and performance.

To address these issues, this paper investigates human inspection behavior analytics for quantifying the reliability of structural component-level defect localizations. The contributions of this research include: (1) designing a bridge inspection game that integrates information from the FEM and inspection report to capture inspectors' behaviors and inspection results in various contexts; (2) discovering and explaining inspectors' strategies through process mining; and (3) quantifying bridge element-level defect localization reliability through inspection behaviors using crowdsourcing and ensemble learning.

4. Methodology

This paper proposes a framework for quantifying the confidence of defects identified by bridge inspectors through behavioral process graph analysis from a digital bridge inspection game, as illustrated in Fig. 5. The framework comprises three steps: (1) gamifying bridge inspection for collecting inspector behavioral logs; (2) employing process mining to uncover inspectors' inspection strategies; and (3) using crowdsourcing to quantify defect localization reliability.

4.1. Gamification of bridge inspection from multi-modal data

During a bridge inspection, inspectors are onsite and search for flaws, defects, or potential problem areas that may necessitate maintenance. Obtaining the bridge geometry, defect locations, and related defect images is crucial to ensure the digital bridge inspection closely aligns with real-life situations. Our proposed solution integrates multi-modal data: (1) element positions (X, Y, Z), displacements, and stress from the FEM simulation software Ansys; (2) basic bridge information and possible defect types from practical inspection reports; and (3) bridge geometry information from Revit, as shown in Fig. 6. To accurately represent element-level defects, FEM simulates defective structural elements based on crack locations from inspection reports spanning several years. In this way, inspectors can access various defects in the bridge inspection game, informed by physics-based FEM simulation results.

Inspectors can click on an element to check its stress and displacement in the properties panel (see Fig. 7). When they identify an element likely to have defects, they must complete inspection records, including defect locations (in which section of the bridge, ranging from 0 to 152), structure types (on which element of the box girder: bottom slab, left web, right web, or top slab, labeled as 0–3), damage types (transverse crack, diagonal crack, or unknown), damage severity (good, fair, poor, severe, or unknown), and the defect's cause (torsion, bending, shearing, tension, compression, or unknown) (see Fig. 7). By comparing inspectors' final inspections of the defects with the ground truth simulated by FEM, inspectors' performance in bridge inspection can be evaluated.

4.2. Process discovery and conformance checking

4.2.1. Event log preparation

A process is defined as a series of activities or steps executed by performers to meet specific requirements or objectives (Kouhestani and Nik-Bakht, 2020). Defining the event log structure is essential to represent the process as a standard and suitable data structure. A formal process consists of cases (traces), events, timestamps, activities, related resources, and costs. Each case includes a collection of events, and each event is associated with the execution of an activity (Kouhestani and Nik-Bakht, 2020). The attribute "inspector" represents the event's case because each inspector's bridge inspection is treated as one process. To protect participants' privacy, all personal information has been removed or replaced with a serial number.

Furthermore, activities are the attribute lists "section" (0–152) and "element" (0–3), which represent the sequences of checked bridge segments and elements. The resources are the inspectors who executed the activity. One inspector is responsible for the entire bridge inspection in

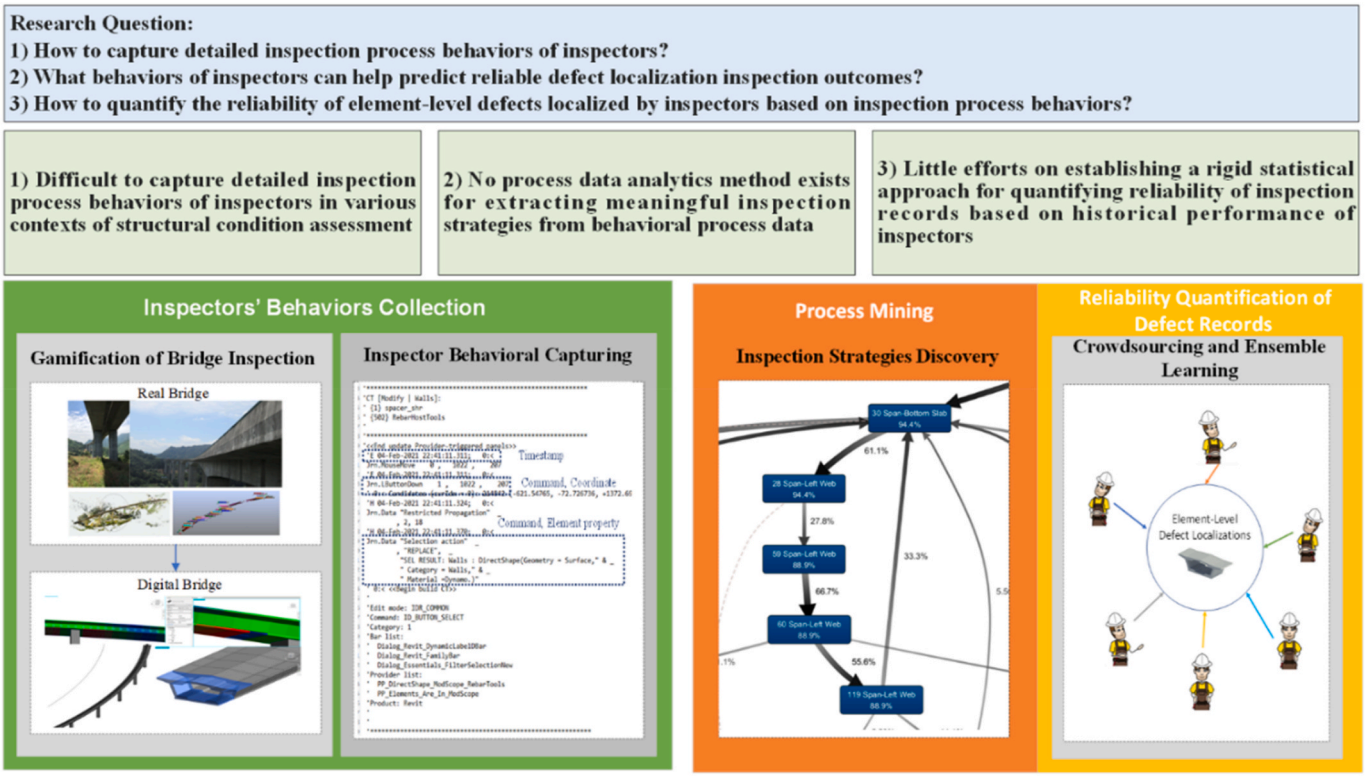


Fig. 5. Framework for quantifying the reliability of defects located by bridge inspectors through behavioral process analysis.

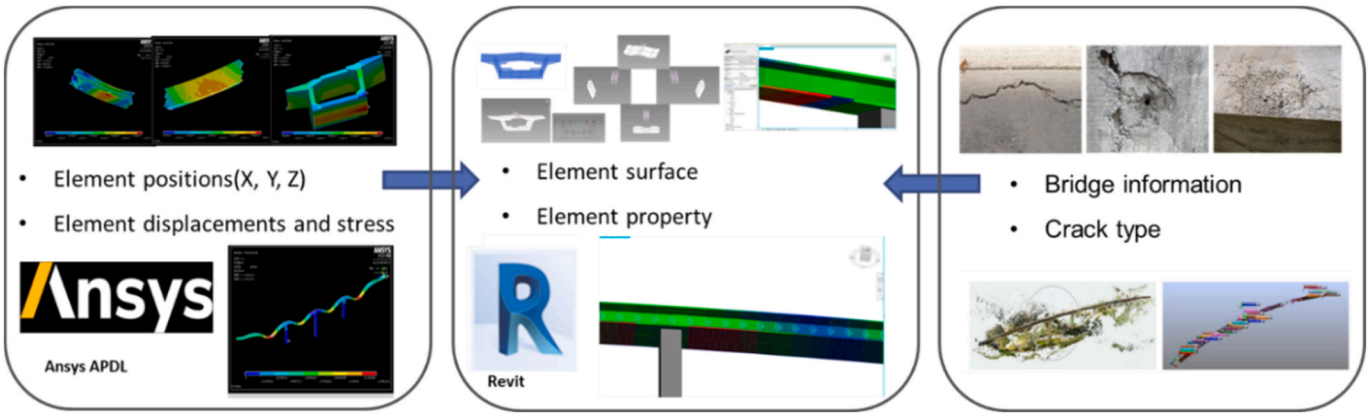


Fig. 6. Multi-model data integration-based bridge inspection gamification.

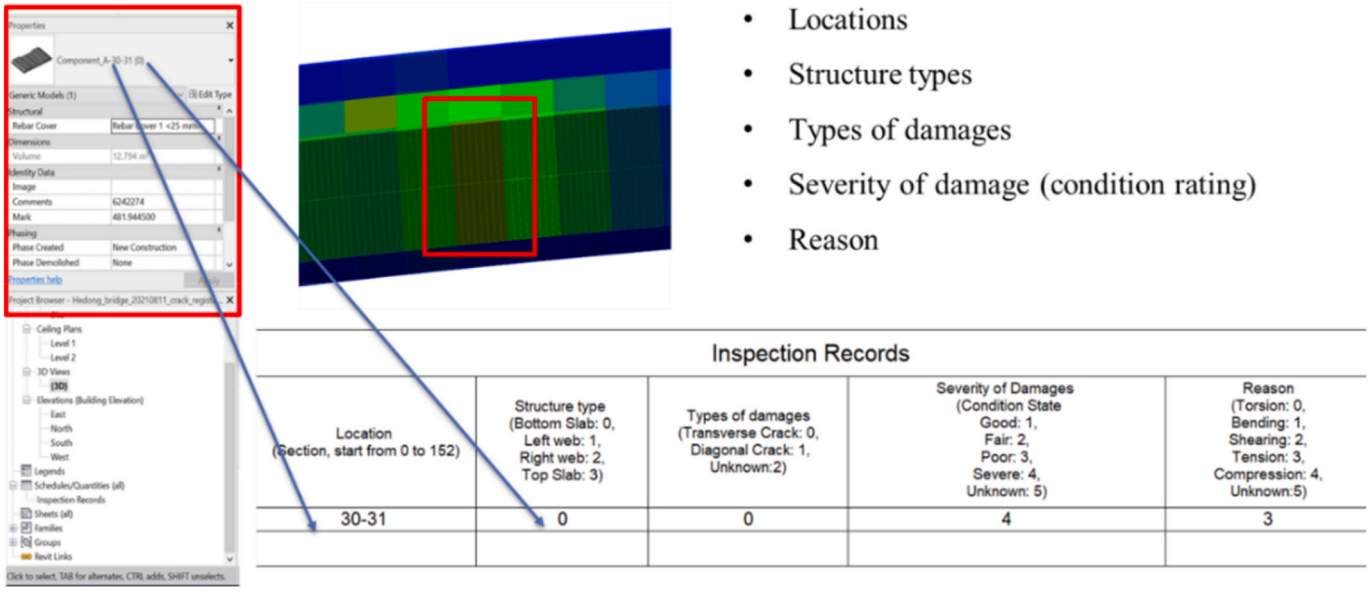


Fig. 7. Digital inspection records.

our tasks. For statistical purposes, the authors use the classification of inspectors as resources. However, resources could include several people in other cases or domains to accomplish the same task sequence. The last column is the timestamp of each activity. This information helps analyze related properties, such as each activity's duration and the waiting time between two activities.

4.2.2. Process model representation and process discovery algorithms

A process model is an abstraction of the complex process recorded in event logs, which can be visualized in different forms to better describe execution sequences and dependencies in a series of activities. The Petri net, initially developed in the late 1960s, is one of the most prominent process modeling languages that combines mathematical formalism and graphical representation, facilitating the display of both concurrency and asynchrony in processes. The Petri net is typically a bipartite graph, where a set of directed arcs connects places in circles and transitions in squares, representing various associations. There are four basic relations: (1) Direct succession: $x > y$ if, for some case, x is directly followed by y; (2) Causality: $x \rightarrow y$ if $x > y$ and not $y > x$; (3) Parallel: $x||y$ if $x > y$ and $y > x$; (4) Choice: $x\#y$ if not $x > y$ and not $y > x$. Five basic process patterns can be discovered from event logs, as shown in Fig. 8, including sequence pattern, XOR-split pattern, XOR-join pattern, AND-split pattern, and AND-join pattern (van der Aalst, 2010): (a) Sequence pattern: $a \rightarrow b$, activity b occurs immediately after activity a; (b) XOR-split pattern: $a \rightarrow b$, $a \rightarrow c$, and $b\#c$, activity b or c occurs after activity a (takes one outgoing branch); (c) XOR-join pattern: $b \rightarrow d$, $c \rightarrow d$, and $b\#c$, activity d occurs after either activity b or c occurs (proceeds when one incoming branch is completed); (d) AND-split pattern: $a \rightarrow b$, $a \rightarrow c$, and $b||c$, activities b and c occur after activity a occurs (takes all outgoing branches); (e) AND-join pattern: $b \rightarrow d$, $c \rightarrow d$, and $b||c$, activity d occurs after both activities b and c occur (proceeds when all incoming branches are completed).

4.2.3. Conformance checking for quantity reliability of individual inspector

The process discovery algorithms construct the process model without any prior information. It is essential to evaluate the generated process model's performance in describing the observed behaviors in the event logs. Conformance checking compares an event log and the mined process model to quantify the differences between the discovered process model and the observed behavior (Buijs et al., 2012). Replay fitness

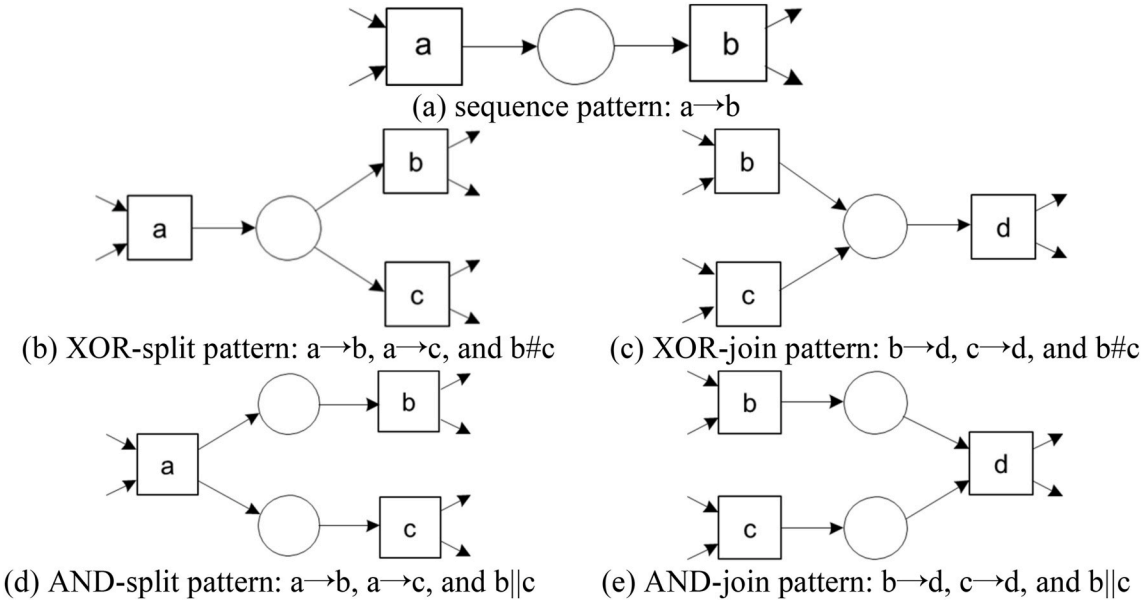


Fig. 8. Illustration of the process model.

quantifies the extent to which the discovered model can accurately reproduce and describe the cases recorded in the event log (Buijs et al., 2012), defined by an alignment-based calculation in Eq. (1) (Buijs et al., 2012; Van der Aalst et al., 2012). The generated fitness can be used to quantify the extent to which an individual inspector follows a behavioral pattern represented by a process model. Fitness with a process model indicating processes learning to reliability defect localization results can serve as an indicator of an inspector's reliability. In other words, it measures the similarity between the discovered process model and the replayed event logs. While aligning events to the process model, a cost is assigned when events are skipped, or activities are inserted without expectation. If all cases from logs are fully reproduced, a perfect fitness score closer to 1 is obtained. Conversely, a fitness of 0 signifies that the process model fails to replay traces in the log.

$$\textit{Inspection reliability index}, \textit{fitness}(L, M) = 1 - \frac{\textit{fcost}(L, M)}{\mathrm{move}_{\mathrm{L}}(\mathrm{L}) + \mathrm{move}_{M}(\mathrm{M})} \quad (1)$$

Where the $\textit{fcost}(L, M)$ function, also known as alignment cost, is a measure of the extent to which an event log (L) aligns with a process model (M). It quantifies the deviations between the actual process captured in the event log and the ideal process defined by the model. A lower fitness cost indicates better alignment between the log and the model. For example, if $\textit{fcost}(L, M) = 0$, it means that model M can perfectly replay the log L. For the denominator, it stands for the maximal possible cost, where $\mathrm{move}_{\mathrm{L}}(\mathrm{L})$ is the cost of moving through the whole logs instead of the model, and $\mathrm{move}_{M}(\mathrm{M})$ is the cost of making moves on the model.

To calculate fitness using $\textit{fcost}(L, M)$, $\mathrm{move}_{\mathrm{L}}(\mathrm{L})$, and $\mathrm{move}_{M}(\mathrm{M})$), follow these steps: (1) Find the optimal alignments between the event log traces and the process model paths using an algorithm like the A* algorithm or other search heuristics. (2) For each trace in the event log, calculate the alignment cost as the sum of the costs of the deviations in the optimal alignment. Commonly, a cost of 1 is assigned to each move in the log (a recorded event not allowed by the model) and each move in the model (an allowed transition not observed in the log). (3) Calculate $\mathrm{move}_{\mathrm{L}}(\mathrm{L})$ as the total number of moves in the log, and $\mathrm{move}_{M}(\mathrm{M})$ as the total number of moves in the model. (4) Calculate the total alignment cost as the sum of the alignment costs for all traces in the event log.

For example, consider the following event log and process model: Event Log (Trace 1): A - > C - > B - > D; Process Model: A - > B - > C - > D. Assuming we have found the optimal alignments, we can calculate fitness using $\textit{fcost}(L, M)$, $\mathrm{move}_{\mathrm{L}}(\mathrm{L})$, and $\mathrm{move}_{M}(\mathrm{M})$) as follows: 1) Optimal alignments: Trace : (A, A) -> (C, ⊥) -> (B, ⊥) -> (⊥, B) -> (⊥, C) -> (D, D) (4 deviations); 2) Calculate $\mathrm{move}_{\mathrm{L}}(\mathrm{L}) = 2$ (C and B in Trace) and $\mathrm{move}_{M}(\mathrm{M}) = 2$ (B and C in Trace); 3) Calculate the total alignment cost: 4; 4) Calculate the fitness: 1 - (4/(2 + 2)) = 1 - (4/4) = 1 - 1 = 0.

In this example, the fitness is 0, indicating a low level of conformance between the event log and the process model. The $\mathrm{move}_{\mathrm{L}}(\mathrm{L})$ and $\mathrm{move}_{M}(\mathrm{M})$ functions both have a value of 2, representing the deviations in the log and the model, respectively. These values can be used to identify areas for improvement in the process model and the event log.

4.3. Quantifying the reliability of defect localization records

4.3.1. Quantifying the performances of inspectors and inspection teams

Bridge inspection is a type of fault detection process. Inspectors aim to find the defects among all the elements. Compared to normal elements, the numbers of abnormal elements are scarce. Therefore, the commonly used metrics in anomaly detection are also suitable for evaluating the inspectors' performances. In addition, as shown in Fig. 13, by comparing the inspectors' final inspection of the defects and the ground truth of defects simulated by FEM, inspectors' performances in the bridge inspection could be evaluated. In this research, elements with defects are known as the positive class, whereas defect-free elements are considered negative. The well-known performance scores are used in the evaluation process: defect location accuracy for both individual inspectors and inspection teams (*precision, recall, F1 score, and false alarm*) and defect location reliability for the inspection teams (*Fleiss' kappa*) (Falotico and Quatto, 2015; Rücker et al., 2012). The calculation formula is shown below.

Defect Location Performances:

$$\textit{Recall (defect detection rate)} = \frac{\textit{True Positive}}{(\textit{True Positive} + \textit{False Negative})} \quad (2)$$

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)} \tag{3}$$

$$F1 = \frac{2 * Recall * Precision}{(Precision + Recall)} \tag{4}$$

$$\text{False Alarm} = \frac{False\ positive}{Negative} \tag{5}$$

Where True positive: fault samples correctly diagnosed as a fault; True negative: normal samples correctly diagnosed as normal; False positive: fault samples incorrectly identified as normal; False negative: normal samples incorrectly identified as a fault. Furthermore, precision measures the proportion of positive test results that are true positives, also referred to as positive predictive value. Recall measures the proportion of actual failures which are correctly identified. The F1 score is the harmonic mean of precision and recall.

Defect Location Reliability:

$$Fleiss'\ kappa\ \kappa = \frac{\overline{p} - \overline{p}_e}{1 - \overline{p}_e} \tag{6}$$

Fleiss' kappa (Fleiss, 1971) is a statistical measure for assessing the reliability of agreement between a fixed number of raters when assigning categorical ratings to a number of items or classifying items. Where $1 - \overline{p}_e$ gives the degree of agreement that is attainable above chance, and $\overline{p} - \overline{p}_e$ gives the degree of the agreement achieved above chance. *Fleiss' kappa* $\kappa < 0$ stands for poor agreement, $0.01 < \kappa < 0.20$ stands for slight agreement, $0.21 < \kappa < 0.40$ stands for fair agreement, $0.41 < \kappa < 0.60$ stands for moderate agreement, $0.61 < \kappa < 0.80$ stands for substantial agreement, $0.81 < \kappa < 1$ stands for almost perfect agreement.

4.3.2. Problem definition of crowdsourcing for defect localization

The bridge defect localization evaluation could also be considered as crowdsourcing from multiple inspectors' subjective judgments and related historical inspection records. During the bridge inspection, groups of inspectors check the bridge elements to find the defects and decide their functional states based on the visual inspection. After the onsite inspection, the final defect localizations of bridge elements consider different defect localizations contributed by all inspectors, indicating that all inspectors have the same experience and abilities.

A crowdsourced labeled dataset D_1 (defect localizations) consists of M inspectors, which is denoted by $D_1 = \{\langle x_{1i}, y_{1i}, l_{1i} \rangle\}_{i=1}^{I}$. x_i is the behavior log inspectors i and y_i is the true localizations of defect. l_i is the label of inspector i. A process model $P(D_1) = P(\{x_{1i}, y_{1i}\}_{i=1}^{I})$ is generated from the labeled dataset D_1.

A crowdsourced unlabeled dataset D_2 (defect localizations) consists of M inspectors, which is denoted by $D_2 = \{\langle x_{2i}, y_{2i}, r_i \rangle\}_{i=1}^{I}$. x_i is the behavior log inspectors i and y_i is the true localization of defect. r_i is the reliability index of inspector i, is generated from the built process model $P(x_{2i}, y_{2i}) = r_i$.

The goal is to learn a hypothesis $h(x)$ that can minimize the defect localization error:

$$\varepsilon(h(x)) = Pr\left((h(x_2, r) \neq y\right) \tag{7}$$

4.3.3. Ensemble learning framework with crowdsourcing

In this study, according to the inspectors' performance finding the defects discussed in subsection 4.3.1, inspectors could be classified into different skill levels. Next, according to the process model built by inspectors during the mining process, we could understand the inspection strategies of inspectors of different skill levels. Furthermore, the process model built from the high-performance inspectors could be chosen as the classifier to get the "inspection reliability index (fitness)" through conformance checking, as discussed in subsection 4.2.3, to quantify to what level an individual inspector follows a behavioral pattern represented by a process model. The generated inspection reliability index (fitness) stands for the similarity of the behavior between this inspector and the crowded high-performance inspector. Then we could use the generated inspection reliability index as weights to calibrate the final defect localizations of bridge elements. Therefore, we proposed an algorithm with ensemble learning from crowdsourcing, as shown in Fig. 9 and Algorithm 1.

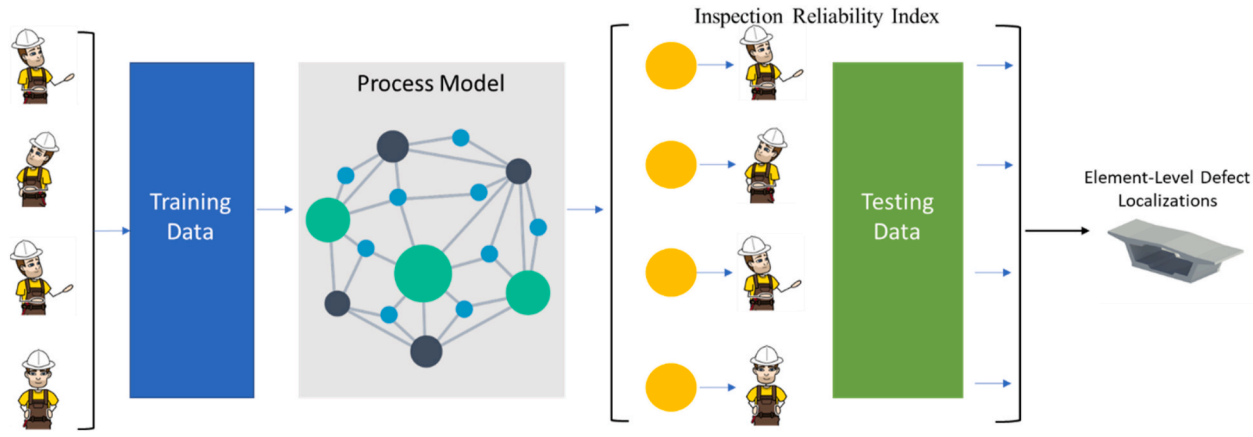


Fig. 9. Illustration of ensemble learning framework with crowdsourcing.

Algorithm 1. Ensemble Learning from Crowdsourcing

Algorithm 1. Ensemble Learning from Crowdsourcing

Input:

The crowdsourced labeled training dataset D_1 $\{< x_{1i}, y_{1i}, r_{1i} >\}_{i=1}^{N}$ from one group of inspector log

The crowdsourced unlabeled testing dataset D_2 $\{< x_{2i}, y_{2i}, r_{2i} >\}_{i=1}^{M}$ from another group of inspector log

The number of base classifiers (inspectors) M

Output:

An ensemble classifier H(x) for calibration of the defect localization

1. Load the inspector log dataset D_1 with labels
2. Create process model P from log dataset D
3. Load the inspector log dataset D_2 without labels
4. Initial M base classifiers (inspectors) h_i
5. **For** $i = 1$ *to* M **do**
6. Generate inspection reliability index r_i through conformance checking of the process model P for the inspector log dataset D_2
7. Assign the generated inspection reliability index r_i to base classifiers (inspectors) $h_i(ri)$
8. **End for**
9. Aggregate M base classifiers (inspectors) into a strong classifier, i.e., $H(x) = F(h1(x_{21}, r_1), h2(r_2), \ldots, hM(r_M))$
10. **Return** $H(x)$

5. Experimental results and discussion

To validate the proposed framework for collecting bridge inspectors' behaviors and calibrating element-level defect localizations through weighted inspection behavior analysis, this study gathers and analyzes BIM event logs from 96 graduate students with basic structural engineering knowledge from the School of Architecture and Department of Civil and Environmental Engineering at Carnegie Mellon University in 2021 (48 students) and Fall 2022 (48 students). In this validation, the authors use log data from Fall (2021) to build and test the process model with data from Fall (2022).

5.1. Gamification of bridge inspection for behaviours collection

This paper employs a typical continuous rigid frame bridge (CRFB) as a case study. This bridge has a total length of 1010 m, including a six-span main bridge of 612 m (66 m + 4*120 m + 66 m), as illustrated in Fig. 10. The main bridge uses prestressed concrete continuous rigid frames with high hollow thin-walled piers (Sun et al., 2019). During the bridge inspection, dividing a bridge into several segments is common, allowing inspectors to examine sections sequentially. This study divides the bridge into 153 segments, numbered from 0 to 152. In each section, a box girder forms an enclosed tube or hollow box-like structure with multiple walls. The box girder is divided into four elements, including

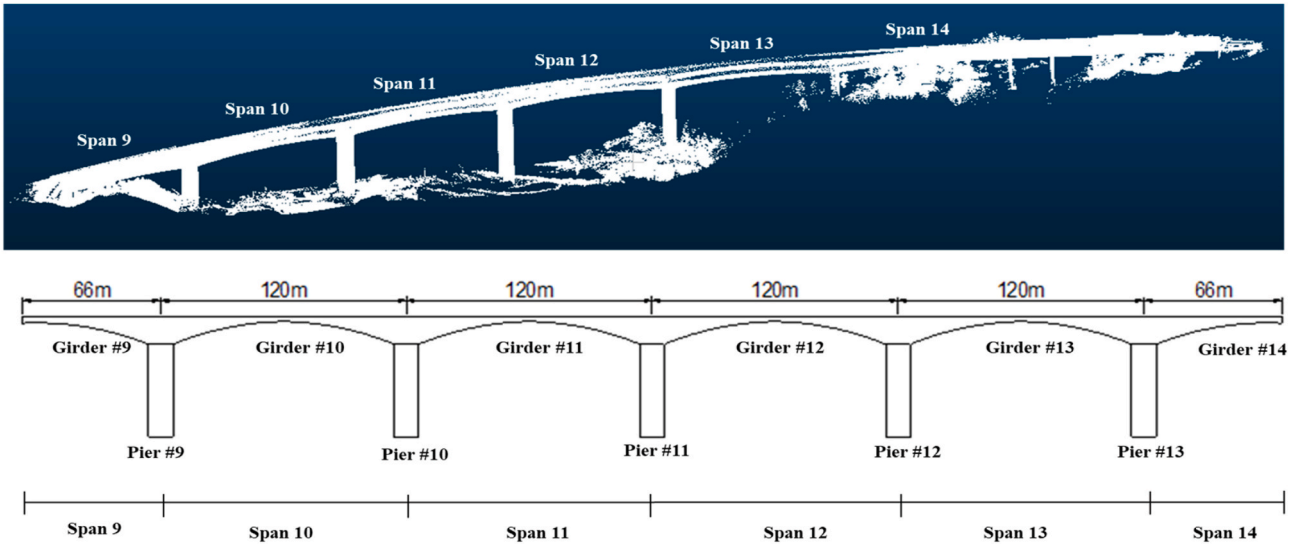


Fig. 10. Continuous rigid frame bridge with four middle spans and two side spans.

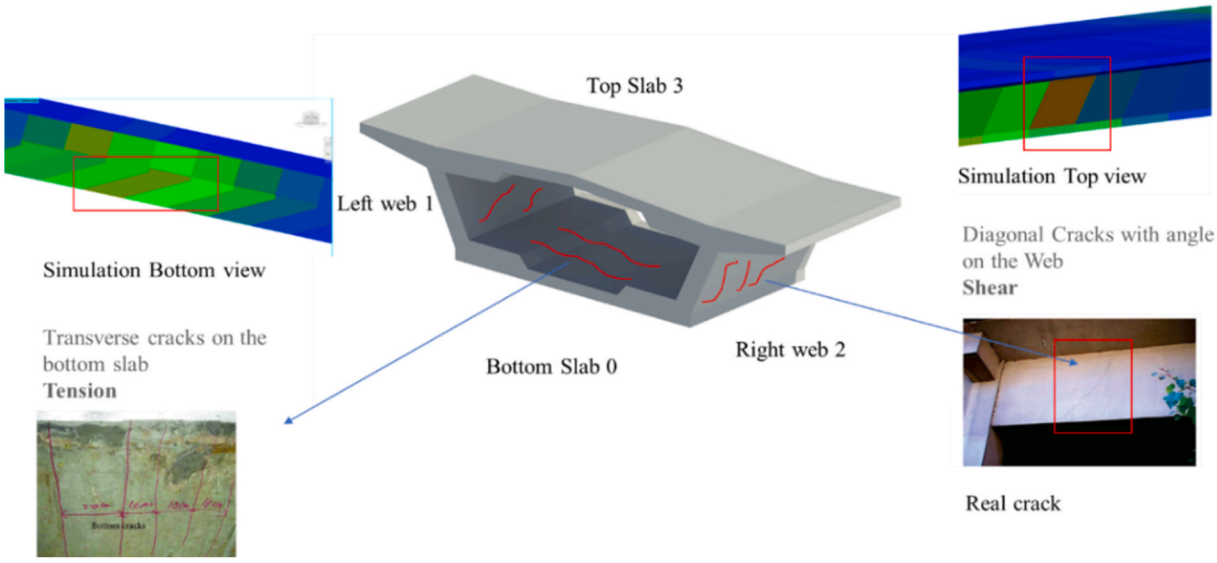


Fig. 11. Possible defects in elements of box girder.

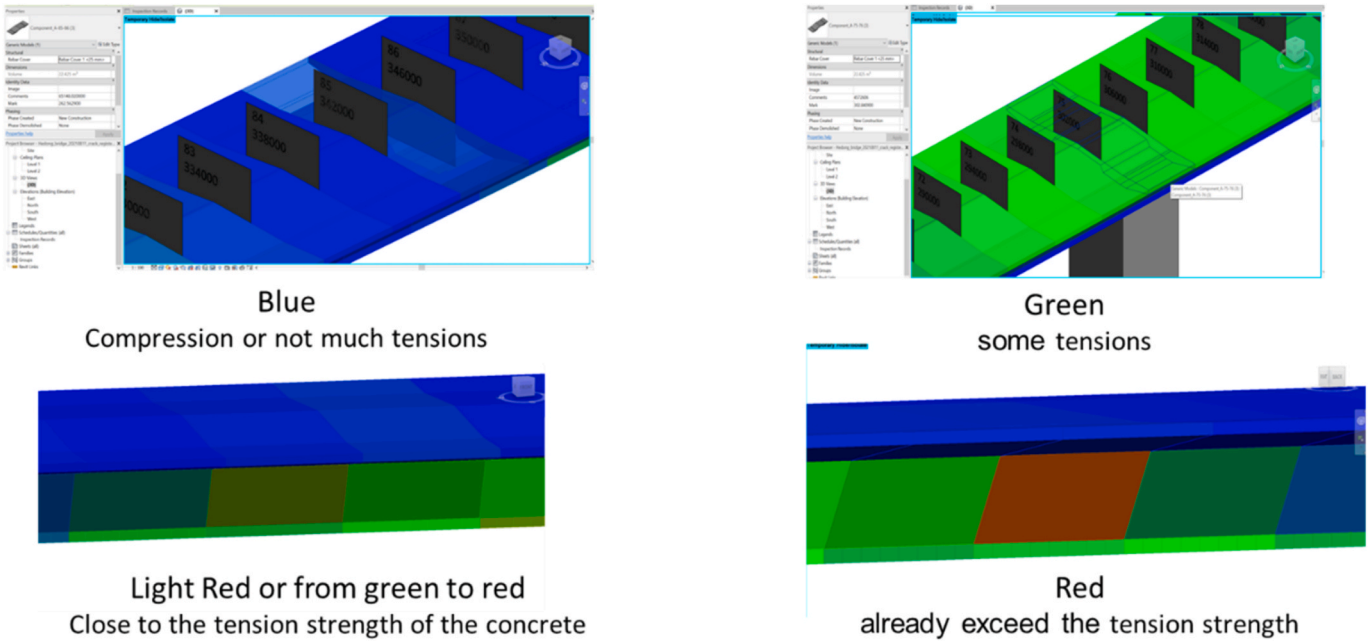


Fig. 12. Colors represent the magnitude of stress

bottom slab 0, left web 1, right web 2, and top slab 3, enabling inspectors to click each element to check the functional status, as shown in Fig. 11. We created a FEM to simulate defect developments on the box girders' slabs and webs in the middle of spans 10, 11, and 13 of CRFBs through stiffness reduction of those elements in the FEM.

After simulating defective structural elements using FEM, element stresses, and displacements are exported from Ansys through parametric design language into Revit to represent functional states and possible defects, as depicted in Figs. 11 and 12. For example, transverse cracks often appear in the bottom slab of the box girder when bearing tension exceeds ultimate strength. Diagonal cracks with angles typically occur on the box girder's web when bearing shear surpasses ultimate strength. Different colors, ranging from blue and green to red, represent stress magnitudes, providing inspectors with a visual sense. Blue indicates good functional states with compression or limited tension; green represents appropriate functional states with some tension; light red or colors between green and red signify poor functional states with stress near ultimate strength; red denotes severe functional states with stress exceeding ultimate strength.

During the inspection, inspector behaviors are recorded as BIM event logs, capturing timestamps, mouse movement coordinates, mouse clicks, selected structural elements, etc. As illustrated in Fig. 13, the logs show that the inspector clicked the left mouse button to select the top slab element in segments 37–38 of the bridge on June 04, 2022. Thus, we can understand the inspector's strategy via selection sequences, event interval times, and mouse movements through the BIM event logs.

5.2. *Process mining for inspection strategy discovery*

The F1 score is chosen as the primary metric to evaluate inspectors' overall anomaly detection performance. The F1 score is a classification accuracy metric that combines precision and recall, designed to be useful when classifying between unbalanced classes. Based on the F1 scores, inspectors can be classified into three types: low performance

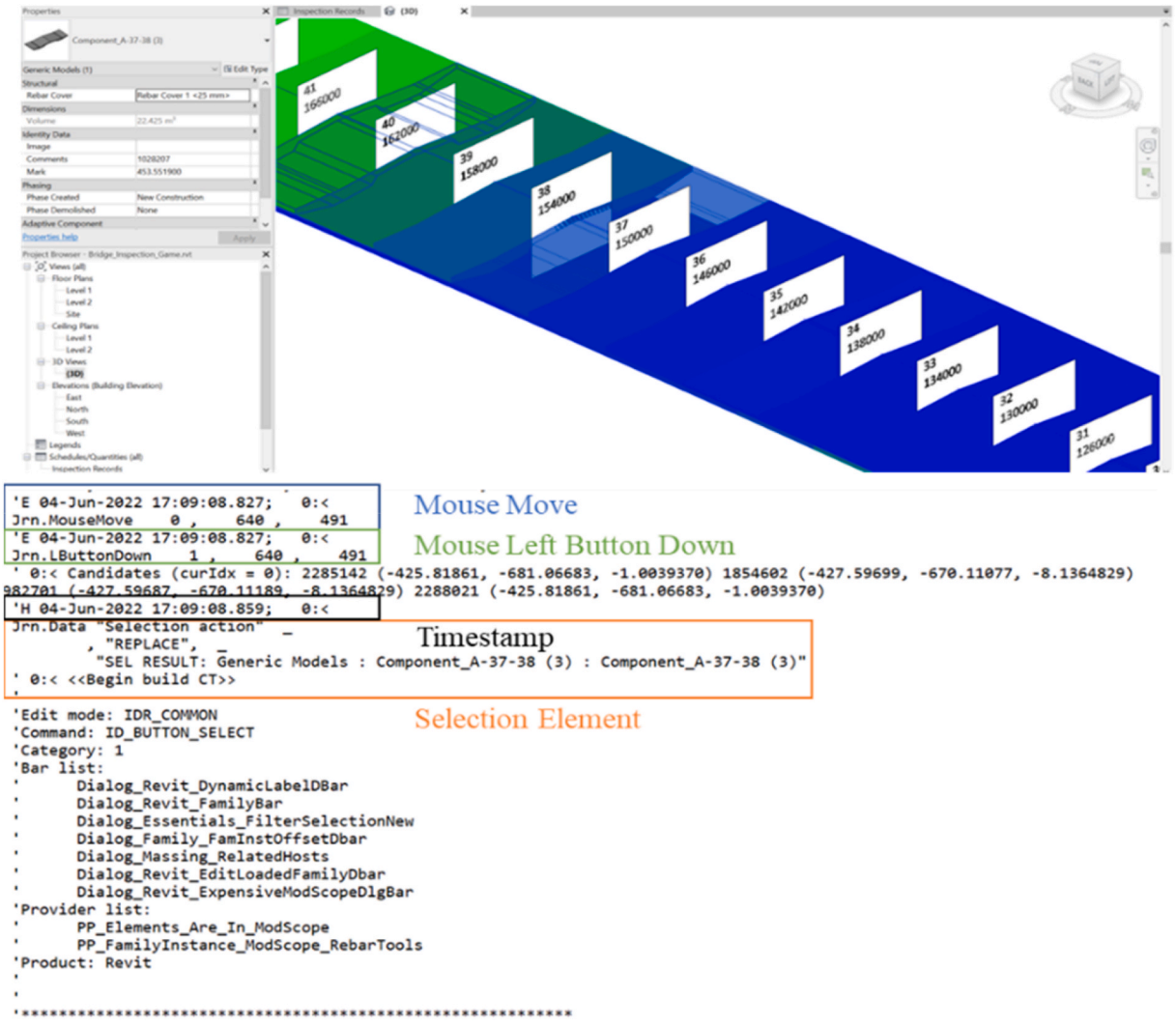


Fig. 13. Revit logs for recording inspection behaviors.

(0–0.5, 18 inspectors), medium performance (0.5–0.85, 46 inspectors), and high performance (0.85–1, 34 inspectors).

Because the inspection includes selections of 153 segments and each section has four elements, the complete process model has hundreds of nodes and complex connections that are difficult for humans to read and interpret the process model. Therefore, the authors simplify the 153 segments into side span 9 (from section 0 to section 16), middle span 10 (from section 16 to section 46), middle span 11 (from section 46 to section 76), middle span 12 (from section 76 to section 106), middle span 13 (from section 106 to section 136) and side span 14 (from section 136 to section 152) to make the Petri net process model easy to understand.

5.2.1. *Process analysis for inspectors with high inspection performances per F1-score*

Fig. 14 displays the search strategies of bridge inspectors with high performance ($F1 > 0.85$), with darker boxes representing more frequent selections of bridge parts by human subjects. Notably, inspectors select right and left webs, top and bottom slabs in span 10, as shown in the red circle of Fig. 14. The right, left webs, and bottom slabs in the middle span 10 receive more selections (163, 136, and 194) compared to top slabs (87). The right and left webs and the bottom slabs are also consistent with defect locations in span 10. Inspectors exhibit similar behavior in spans 11 and 13.

In span 12, without obvious defects, inspectors identify only one location with a high likelihood of defects, as shown in the orange circle. This suggests that high-performing bridge inspectors recognize relationships and stress-transfer mechanisms within spans 10, 11, 12, and 13, leading to high probabilities of jumping between different elements inside spans. Based on this state transition diagram, inspectors appear to focus on the structural similarity between middle spans (10, 11, 12, and 13), with arrows connecting different middle spans to represent transitions from one span to another.

5.2.2. *Process analysis for inspectors with medium inspection performances per F1-score*

The process model for inspectors with medium performance ($0.5 < F1 < 0.85$) displays similar patterns to the Petri net process model of inspectors with high performance, as shown in Fig. 15. Inspectors with medium performance can identify relationships between and within spans. However, as shown in the green circle, they overlook the relationship of elements in span 13, where the right web is isolated from the bottom and left slabs.

Fig. 15 also shows span 11 in two blue circles, representing different elements (slabs or webs) of that cross-section, without clear connections between different spans. In summary, compared to inspectors with high performance, those with medium performance tend to ignore relationships and stress-transfer mechanisms within and between spans 10, 11, 12, and 13.

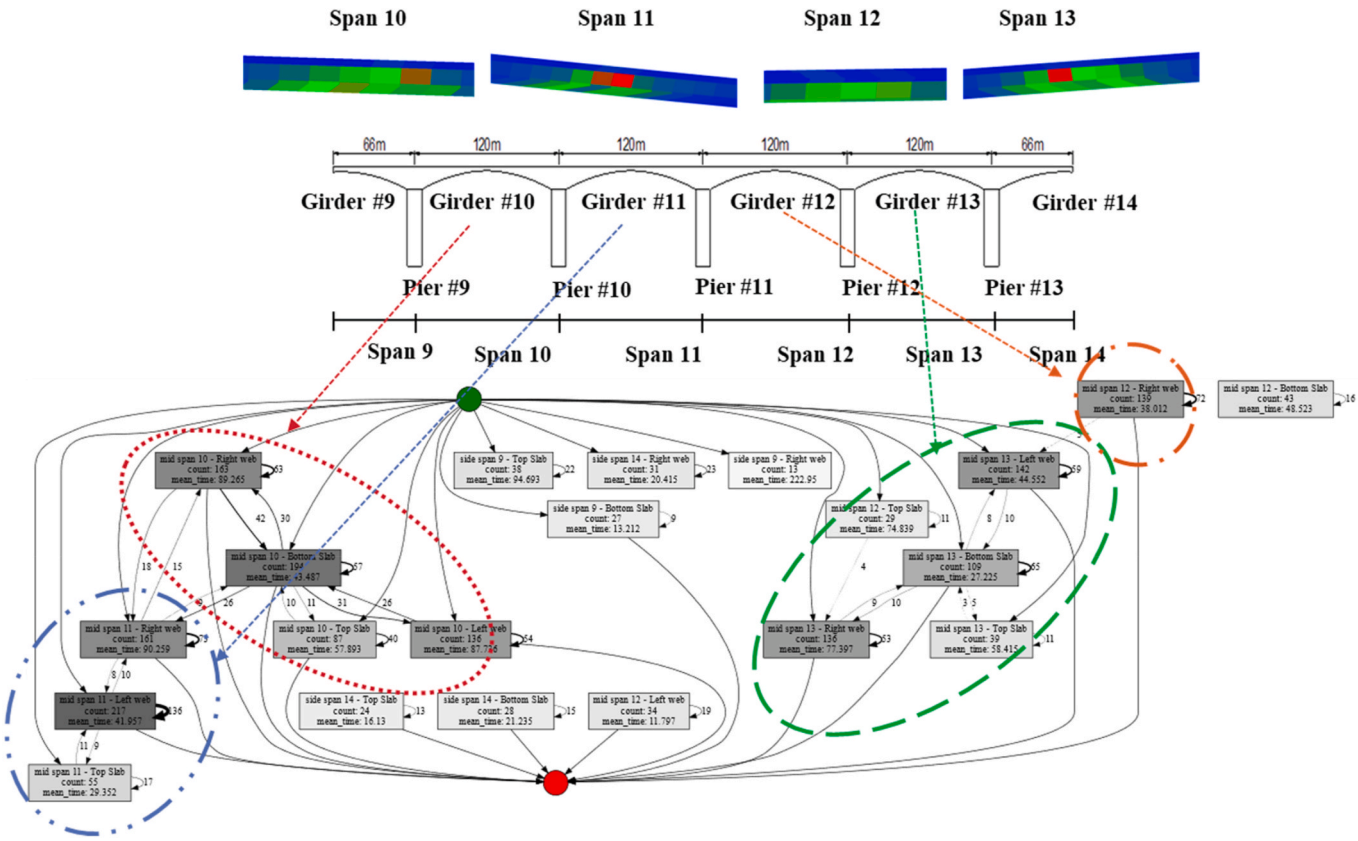


Fig. 14. Process Mining for the inspectors with high performances.

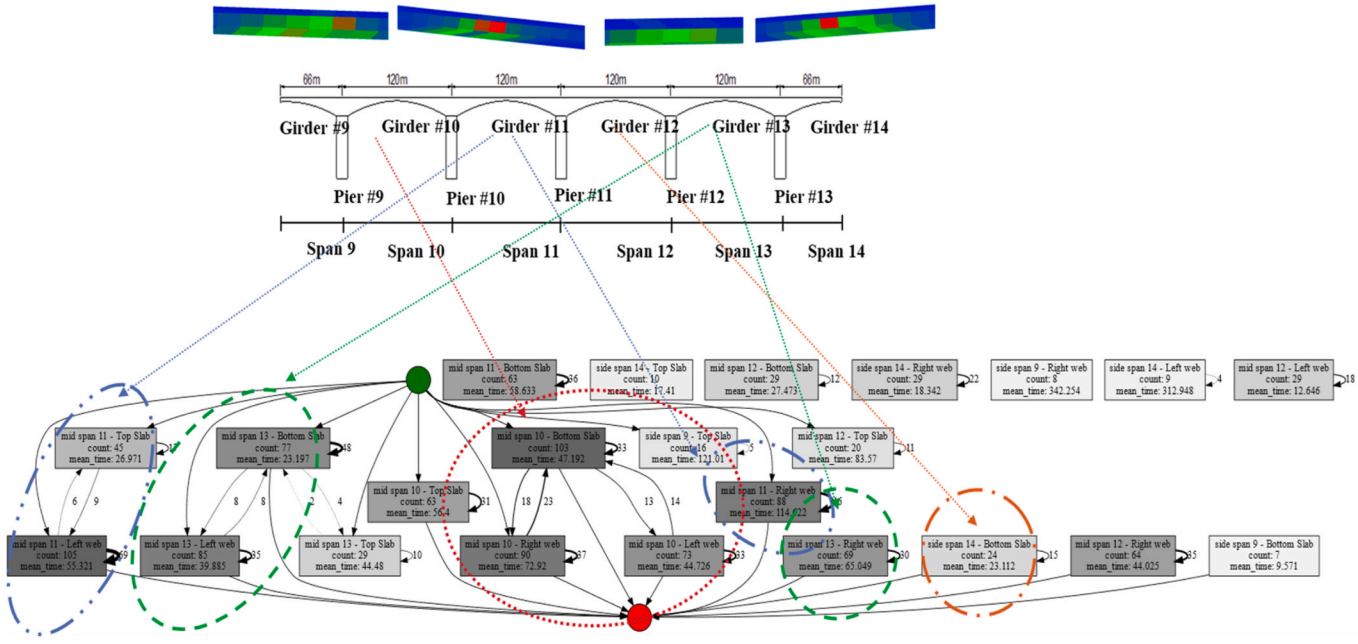


Fig. 15. Process mining for the inspector with medium performances.

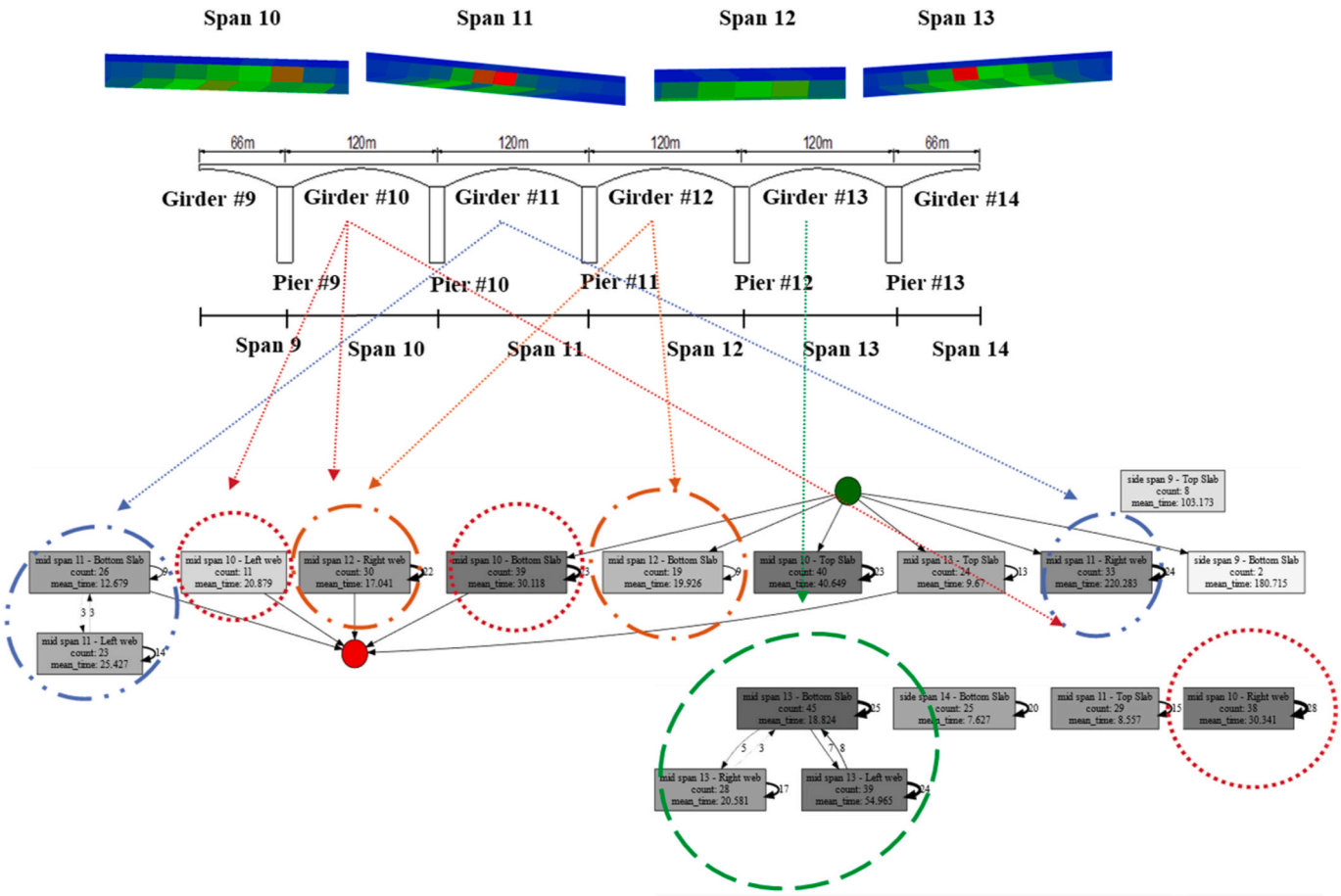


Fig. 16. Process mining for inspectors with low performances.

5.2.3. Process analysis for inspectors with low performances per F1-score

For inspectors with low performance ($F1 < 0.5$), the process model reveals that different elements in the same span are separated into several parts, as shown in different circles in Fig. 16, for example, from one another. The process model indicates that low-performance inspectors may employ random search strategies and disregard the relationships and stress-transfer mechanisms within and between spans 10, 11, 12, and 13.

5.3. Quantifying the reliability of reported locations of bridge element cracks through weighted inspection behaviour analysis

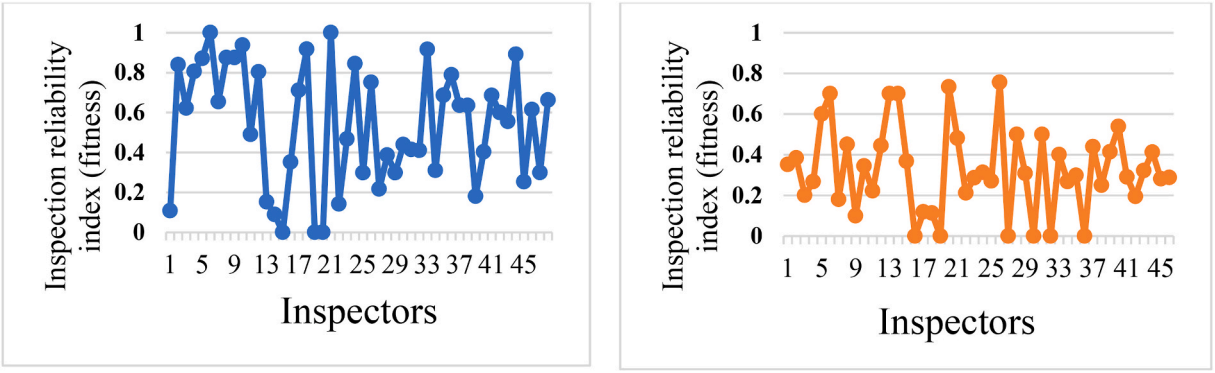
The fundamental concept is that bridge defect locations can be considered a crowdsourcing process stemming from multiple inspectors' subjective judgments and related inspection records. During bridge inspection, groups of inspectors assess bridge elements to determine defect locations based on visual inspection. After the onsite inspection, the reliability and performance of the final defect locations can be calculated using two methods. One method involves calculating the mean value of different defect locations, assuming all inspectors have the same skill level. As discussed in subsection 4.3, the second method involves evaluating inspectors' inspection behavior through a process graph model to generate an inspection reliability index. This generated inspection reliability index can be used as a weight to account for inspectors' skill levels when voting for defect locations of bridge elements, thereby improving reliability.

5.3.1. Generation of inspection reliability index (fitness)

The inspectors' previous behaviors are recorded in a process model. It is a graph model with directional arrows between actions and events in inspection processes. The similarity between the current and previous ones is calculated based on conformance checking through the token-based replay, as discussed in subsection 4.2.3. Based on the process model, we can understand the inspection strategies of inspectors with different skill levels. Furthermore, the process model built from high- and medium-performance inspectors can be selected as the classifier to obtain fitness through conformance checking using token-based replay, as shown in Fig. 17. Since we collected data from 96 graduate students at Carnegie Mellon University in the fall of 2021 and 2022, we used the BIM event logs from high- and medium-performance inspectors in the fall of 2021 to build the process model. We then evaluated the model on data from the fall of 2021 (including training data), as shown in Fig. 17 (a), and data from the fall of 2022, as shown in Fig. 17(b). The generated fitness represents the similarity of behaviors between a given inspector and the collective high- or medium-performance inspectors. We find that the rebuilt process model from the data of inspectors with high and medium performance can classify new inspectors' behaviors and generate the inspection reliability index (fitness).

5.3.2. Reliability quantification of localized defect records

The authors compare average and calibrated defect locations to evaluate whether the proposed weighted approach can improve the reliability of element-level defect locations voted on by multiple inspectors. Fleiss' kappa is an index that assesses the reliability of agreement between a fixed number of inspectors when assigning categorical ratings to several items or classifying items (Fleiss, 1971; Falotico and Quatto, 2015). In this study, 96 participants are required to classify 612 bridge elements (153 segments, with four elements in each segment) as



(a) Fitness of process model on data from 2021 fall (including training data of inspectors with high and medium performance)

(b) Fitness of process model on test data from 2022 fall

Fig. 17. Performances of the built process model.

normal or defective. Fleiss' kappa was applied to analyze the variations between multiple inspectors' defect locations. The reliability of defect locations improved from 0.343 (with a 95% confidence interval (CI) of 0.341–0.345) to 0.468 (95% CI 0.466–0.469). The p-values of both tests are less than 0.0005, indicating statistical significance (see Table 1). The experiment results show that the overall reliability of defect locations changes from fair agreement to moderate agreement.

5.3.3. Performance quantification of localized defect records

To evaluate whether the proposed framework with ensemble learning from crowdsourcing, as discussed in subsection 4.3.3, can improve the accuracy of element-level defect locations, the authors compare average and calibrated defect locations. Table 2 and Fig. 17 show that the weighted average can significantly calibrate the element-level defect locations to the ground truth.

An average recall (fault detection rate) of 90.72% for ten defects was achieved for element-level defect locations by aggregating multiple inspectors with ensemble learning through the process model compared to the 82.61% recall (fault detection rate) without using their inspection reliability indices as weights (see Table 3). We also compared the false alarms of ten selected locations to illustrate that the proposed approach could identify false alarms and reduce the probabilities for those erroneously reported defects by some inspectors, as shown in Table 2 and Fig. 18. An average of 13.48% false alarms for ten defects was achieved for element-level defect locations by aggregating multiple inspectors with ensemble learning through the process model compared to the 17.83% false alarm rate without using their inspection reliability indices as weights (see Table 3).

Moreover, we compared the average and weighted average ensemble learning for all defect localizations in terms of accuracy, precision, recall, F1, and false alarm rate. The results indicate that the weighted

Table 3
Comparison of performances of average and weighted average ensemble learning for all defect localizations.

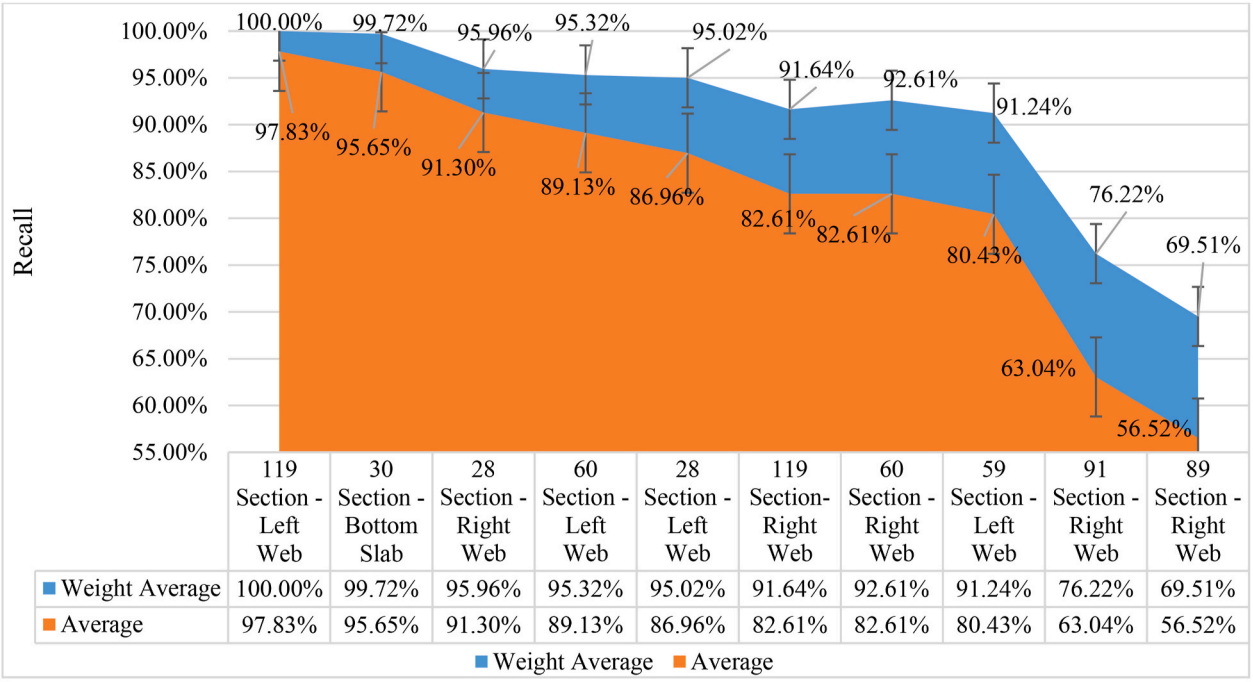
Method	Accuracy	Precision	Recall	F1	False Alarm
Average	0.9342	0.3267	0.8261	0.4682	0.0619
Weight Average	0.9533	0.3594	0.9087	0.5151	0.0451

Table 1
Reliability quantification of the localized defect records.

	Kappa	Asymptotic			Asymptotic 95% Confidence Interval	
		Standard Error	z	Sig.	Lower Bound	Upper Bound
Overall Reliability	0.343	0.001	384.765	0.000	0.341	0.345
Overall reliability with weight	0.468	0.001	547.538	0.000	0.466	0.469

Table 2
Comparison of recall from the average and weighted average ensemble learning.

Defect Location	Recall (Fault Detection Rate)		Defect Location	False Alarm	
	Average	Weight Average		Average	Weight Average
119 Section - Left Web	97.83%	100.00%	61 Section -Left web	21.74%	15.22%
30 Section - Bottom Slab	95.65%	99.72%	120 Section-Right web	19.57%	17.39%
28 Section - Right Web	91.30%	95.96%	31 Section-Bottom Slab	19.57%	13.04%
60 Section - Left Web	89.13%	95.32%	119 Section-Bottom Slab	19.57%	17.39%
28 Section - Left Web	86.96%	95.02%	32 Section-Left web	17.39%	13.04%
119 Section-Right Web	82.61%	91.64%	45 Section-Top Slab	17.39%	15.22%
60 Section - Right Web	82.61%	92.61%	121 Section-Left web	17.39%	10.87%
59 Section -Left Web	80.43%	91.24%	117 Section-Bottom Slab	15.22%	10.87%
91 Section - Right Web	63.04%	76.22%	15 Section-Top Slab	15.22%	10.87%
89 Section - Right Web	56.52%	69.51%	16 Section-Top Slab	15.22%	10.87%
Average	82.61%	90.72%		17.83%	13.48%



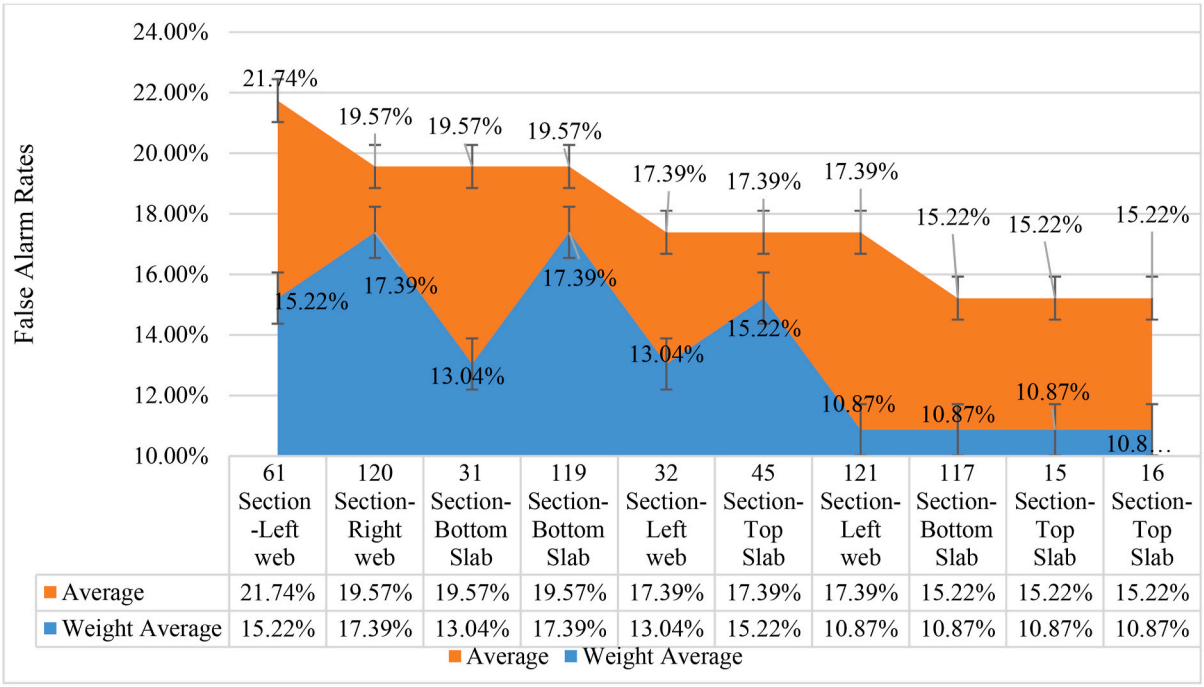


Fig. 18. Comparison of recall and false alarm rates of defect localizations between the average methods and weighted average methods.

average, by aggregating multiple inspectors with ensemble learning through process analysis, achieves better performance than without using their inspection reliability indices as weights, particularly in recall (increasing from 0.8261 to 0.9087) and false alarm rate (decreasing from 0.0619 to 0.0451).

6. Discussion

Several technical and scientific considerations could influence the technical feasibility and reliability of the proposed method in bridge inspection practice. These considerations include 1) differences between the mouse clicking and eye tracking methods for observation behavior tracking; 2) pros and cons of different process data analysis methods; 3) explanation of the discovered strategies through the process mining; 4) consideration about using the behavioral data collected in the virtual environment for calibrating the defect localization records obtained in physical environments. The following paragraphs detail these considerations and their implications for future studies and field implementations.

The first consideration is observing behavior tracking by mouse clicking and eye tracking. In most situations, inspectors click on elements of interest, enabling the BIM log to record click time, selected elements, and sequences, representing inspectors' behaviors. However, in some cases, inspectors notice elements without clicking on them,

leading to missed capture of essential behaviors. Several ways to overcome this problem include using eye-tracking equipment or integrating mouse movement and rotation data for better behavior understanding.

The second consideration is the pros and cons of different process data analysis methods. Process mining analyzes event logs and extracts data from sources like databases and spreadsheets. Recently, deep learning-based models like bidirectional encoder representations from transformers (BERT), a deep learning-based natural language model, are designed for unstructured data like natural language processing tasks. While it is possible to adapt BERT for other applications, such as process mining, it may not be the most efficient or effective approach for processing raw event logs and observation histories. Such event logs and observation histories' semantic information are rather limited compared with words in natural language. BERT could work with a pre-processing step that interprets the raw logs into more meaningful units of human actions and attributes of observed objects for capturing repetitive inspection strategies. However, designing a process ontology to help with that pre-processing is out of the scope of this work. The authors added these discussions to the discussion section to highlight plans for designing process formalism to interpret the event logs for leveraging the BERT approach. In summary, BERT may have some limited use cases in process mining, but it is not a primary method for process mining analysis. In future work, we are considering using the knowledge graph or fine-tuning large language models to represent human knowledge.

The third concern is about the explanation of the discovered strategies through process mining. Inspection strategy discovery methods allow explaining and reusing defect search strategies. However, this paper focuses on revealing repetitive inspection behaviors without explanation and reuse mechanisms. Section 5.2 compares inspection strategies through process analysis of inspectors with varying performance levels (low, medium, and high). In summary, compared to inspectors with low performance, those with high and medium performance tend to adopt a searching strategy to find connections between elements in the same section and across different spans due to their understanding of structural mechanics and stress redistributions. Consequently, we can rebuild the process graph from event logs of high and medium-performance inspectors to identify essential activities and searching strategies. This method can be applied to other scenarios, demonstrating its strong versatility. Through the generated inspection reliability index (individual inspector reliability) by process analysis, the reliability of entire inspector groups for defect location can be improved using the proposed ensemble learning framework in Section 5.3. Additionally, the recall and false alarm rates for defect locations with a high degree of disagreement can significantly improve, as demonstrated by 91 Section - Right Web and 89 Section - Right Web, proving the effectiveness of the proposed framework.

The fourth consideration involves understanding the differences between virtual and real-world inspections as they impact the tool's usefulness in actual projects. Key differences include (1) Environmental factors: Virtual environments lack complex factors like weather conditions, temperature fluctuations, and wind loads that affect real-world inspections. (2) Structural complexity: Virtual models may not capture real-world bridge structures' complexity, affecting tool performance. (3) Sensor and interaction limitations: Virtual environments provide limited sensory input compared to real-world scenarios where inspectors use multiple senses to gather information. (4) Consequences and risk: Virtual environments assume easy accessibility, while real-world inspections involve risks and consequences that affect inspector behavior. (5) Human factors: Virtual environments may not account for factors like fatigue and physical fitness, affecting inspectors' real-world performance.

7. Conclusion and future work

This study presents a bridge inspection game utilizing multi-data sources (BIM, FEM, and inspection reports) to address the challenges of capturing detailed inspection process behaviors of inspectors. By employing process mining for inspection strategy discovery, inspectors' behaviors are represented and evaluated using a process graph as inspection reliability indexes. The generated reliability serves as the crowdsourcing weight in ensemble learning to quantify and calibrate the final crack location results.

We discovered that the BIM log captures inspectors' detailed inspection behaviors, and search strategies focusing on connections between elements within the same section and across different spans typically yield better defect locations. The proposed approach achieves an average 90.72% defect detection rate (recall), compared to the 82.61% rate obtained by aggregating defect detection results of multiple inspectors without behavioral assessment. Inspection reliability analysis of inspectors and inspection teams indicates that using behavioral analysis to weight defect detection results enhances the overall inspection team reliability, improving from "fair agreement" (0.343) to "moderate agreement" (0.69), according to Fleiss' kappa κ. Thus, the proposed framework can increase the accuracy and reliability of defect locations.

However, there are some limitations to our work. First, certain vague defects improve with new approaches, while others do not. Causal analysis and sensitivity analysis may help identify underlying features and reasons for this scenario. Second, future work should incorporate more detailed task-level inspection actions, such as eye tracking, mouse movement, and rotation, to better understand inspectors' search policies.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

This material is based on work supported by the U.S. National Science Foundation (NSF) CAREER under Grant No. 1454654 and the U.S. National Science Foundation (NSF) Convergence under Grant No. 1937115. The authors gratefully acknowledged the support.

References

Al Hattab, M., Hamzeh, F., 2018. Simulating the dynamics of social agents and information flows in BIM-based design. Autom. Construct. 92, 1–22. https://doi.org/10.1016/j.autcon.2018.03.024.

Ball, G., Siemsen, E., Shah, R., 2017. Do plant inspections predict future quality? The role of investigator experience. Manuf. Serv. Oper. Manag. 19 (4), 534–550. https://doi.org/10.1287/msom.2017.0661.

Bogarín, A., Cerezo, R., Romero, C., 2018. A survey on educational process mining. Wiley Interdiscip. Rev. : Data Min. Knowl. Discov. 8 (1), e1230. https://doi.org/10.1002/widm.1230.

Buijs, J.C., Dongen, B.F.v., van Der Aalst, W.M., 2012. On the role of fitness, precision, generalization and simplicity in process discovery. In: OTM Confederated International Conferences" on the Move to Meaningful Internet Systems". Springer. https://doi.org/10.1007/978-3-642-33606-5_19.

Cairns, A.H., Gueni, B., Fhima, M., Cairns, A., David, S., Khelifa, N., 2015. Process mining in the education domain. Int. J. Adv.Intelligent Syst. 8 (1), 219–232. http://personales.upv.es/thinkmind/dl/journals/intsys/intsys_v8_n12_2015/intsys_v8_n12_2015_18.pdf.

Carver, J.C., 2003. The Impact of Background and Experience on Software Inspections. University of Maryland, College Park. ISBN: 0496421638.

Chua, D., Hossain, M.A., 2011. A simulation model to study the impact of early information on design duration and redesign. Int. J. Proj. Manag. 29 (3), 246–257. https://doi.org/10.1016/j.ijproman.2010.02.012.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46. https://doi.org/10.1177/001316446002000104.

Drapeau, R., Chilton, L., Bragg, J., Weld, D., 2016. Microtalk: using argumentation to improve crowdsourcing accuracy. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. https://doi.org/10.1145/3017679.
Drew, T., Võ, M.L.-H., Wolfe, J.M., 2013. The invisible gorilla strikes again: sustained in attentional blindness in expert observers. Psychol. Sci. 24 (9), 1848–1853. https://doi.org/10.1177/0956797613479386.
Falotico, R., Quatto, P., 2015. Fleiss' kappa statistic without paradoxes. Qual. Quantity 49 (2), 463–470. https://doi.org/10.1007/s11135-014-0003-1.
Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. Psychol. Bull. 76 (5), 378. https://doi.org/10.1037/h0031619.
Fletcher, I., Mazzi, M., Nuebling, M., 2011. When coders are reliable: the application of three measures to assess inter-rater reliability/agreement with doctor–patient communication data coded with the VR-CoDES. Patient Educ. Counsel. 82 (3), 341–345. https://doi.org/10.1016/j.pec.2011.01.004.
Gallwey, T., Drury, C.G., 1986. Task complexity in visual inspection. Hum. Factors 28 (5), 595–606. https://doi.org/10.1177/001872088602800509.
Gisev, N., Bell, J.S., Chen, T.F., 2013. Interrater agreement and interrater reliability: key concepts, approaches, and applications. Res. Soc. Adm. Pharm. 9 (3), 330–338. https://doi.org/10.1016/j.sapharm.2012.04.004.
Gomes, A.F.D., Lacerda, A.C. W.G.d., Silva Fialho, J.R.d., 2021. Comparative analysis of process mining algorithms in process discover. In: International Conference on Disruptive Technologies, Tech Ethics and Artificial Intelligence. Springer. https://doi.org/10.1007/978-3-030-87687-6_25.
Günther, C.W., Van Der Aalst, W.M., 2007. Fuzzy mining–adaptive process simplification based on multi-perspective metrics. In: International Conference on Business Process Management. Springer. https://doi.org/10.1007/978-3-540-75183-0_24.
Howe, J., 2006. The rise of crowdsourcing. Wired Mag. 14 (6), 1–4. https://sistemas-humano-computacionais.wdfiles.com/local–files/capitulo%3Aredes-sociais/Howe_The_Rise_of_Crowdsourcing.pdf.
Huang, J., White, R., Buscher, G.. User see, user point: gaze and cursor alignment in web search. https://doi.org/10.1145/2207676.2208591.
Iso, I., Oiml, B., 1995. Geneva, Switzerland. Guide to the Expression of Uncertainty in Measurement, 122, pp. 16–17. http://chapon.arnaud.free.fr/documents/resources/stat/GUM.pdf.
Juran, J.M., De Feo, J.A., 2010. Juran's Quality Handbook: the Complete Guide to Performance Excellence. McGraw-Hill Education, 0071629734.
Kazai, G., 2011. In search of quality in crowdsourcing for search engine evaluation. In: European Conference on Information Retrieval. Springer. https://doi.org/10.1007/978-3-642-20161-5_17.
Kouhestani, S., Nik-Bakht, M., 2020. IFC-based process mining for design authoring. Autom. Construct. 112, 103069 https://doi.org/10.1016/j.autcon.2019.103069.
Laofor, C., Peansupap, V., 2012. Defect detection and quantification system to support subjective visual quality inspection via a digital image processing: a tiling work case study. Autom. Construct. 24, 160–174. https://doi.org/10.1016/j.autcon.2012.02.012.
Leemans, S.J., Fahland, D., Van Der Aalst, W.M., 2013. Discovering block-structured process models from event logs containing infrequent behaviour. In: International Conference on Business Process Management. Springer. https://doi.org/10.1007/978-3-319-06257-0_6.
Liu, Q., Ihler, A., Fisher, J., 2015. Boosting crowdsourcing with expert labels: local vs. global effects. 2015 18th International conference on information Fusion (Fusion). http://ceur-ws.org/Vol-1043/mediaeval2013_submission_56.pdf.IEEE.
Liu, P., Xiong, R., Tang, P., 2021. Mining observation and cognitive behavior process patterns of bridge inspectors, 2021 of Conference. In: ASCE International Conference on Computing in Civil Engineering. https://doi.org/10.1061/9780784483893.075.
Mans, R.S., Schonenberg, M., Song, M., van der Aalst, W.M., Bakker, P.J., 2008. Application of process mining in healthcare–a case study in a dutch hospital. In: International Joint Conference on Biomedical Engineering Systems and Technologies. Springer. https://doi.org/10.1007/978-3-540-92219-3_32.
Megaw, E., Richardson, J., 1979. Eye movements and industrial inspection. Appl. Ergon. 10 (3), 145–154. https://doi.org/10.1016/0003-6870(79)90138–8.
Oyama, S., Baba, Y., Sakurai, Y., Kashima, H., 2013. Accurate integration of crowdsourced labels using workers' self-reported confidence scores. In: Twenty-Third International Joint Conference on Artificial Intelligence. https://doi.org/10.5555/2540128.2540496.
Pan, Y., Zhang, L., 2020a. BIM log mining: exploring design productivity characteristics. Autom. Construct. 109, 102997 https://doi.org/10.1016/j.autcon.2019.102997.
Pan, Y., Zhang, L., 2020b. BIM log mining: learning and predicting design commands. Autom. Construct. 112, 103107 https://doi.org/10.1016/j.autcon.2020.103107.
Pan, Y., Zhang, L., 2021a. Automated process discovery from event logs in BIM construction projects. Autom. Construct. 127, 103713 https://doi.org/10.1016/j.autcon.2021.103713.
Pan, Y., Zhang, L., 2021b. A BIM-data mining integrated digital twin framework for advanced project management. Autom. Construct. 124, 103564 https://doi.org/10.1016/j.autcon.2021.103564.
Pan, Y., Zhang, L., Li, Z., 2020. Mining event logs for knowledge discovery based on adaptive efficient fuzzy Kohonen clustering network. Knowl. Base Syst. 209, 106482 https://doi.org/10.1016/j.knosys.2020.106482.
Phares, B.M., Washer, G.A., Rolander, D.D., Graybeal, B.A., Moore, M., 2004. Routine highway bridge inspection condition documentation accuracy and reliability. J. Bridge Eng. 9 (4), 403–413. https://doi.org/10.1061/(asce)1084-0702(2004)9, 4 (403).
Remenyi, B., Carapetis, J., Stirling, J.W., Ferreira, B., Kumar, K., Lawrenson, J., Marijon, E., Mirabel, M., Mocumbi, A., Mota, C., 2019. Inter-rater and intra-rater reliability and agreement of echocardiographic diagnosis of rheumatic heart disease using the World Heart Federation evidence-based criteria. Heart Asia 11 (2). https://doi.org/10.1136/heartasia-2019-011233.
Rojas, E., Munoz-Gama, J., Sepúlveda, M., Capurro, D., 2016. Process mining in healthcare: a literature review. J. Biomed. Inf. 61, 224–236. https://doi.org/10.1016/j.jbi.2016.04.007.
Rücker, G., Schimek-Jasch, T., Nestle, U., 2012. Measuring inter-observer agreement in contour delineation of medical imaging in a dummy run using Fleiss' kappa. Methods Inf. Med. 51 (6), 489–494. https://doi.org/10.3414/ME12-01-0005.
Salem, H., Helmy, H., 2014. Numerical investigation of collapse of the Minnesota I-35W bridge. Eng. Struct. 59, 635–645. https://doi.org/10.1016/j.engstruct.2013.11.022.
Sun, Z., Tang, P., Shi, Y., Xiong, W., 2019. Visual-semantic alignments for automated interpretation of 3D Imagery data of high-pier bridges. In: Computing in Civil Engineering 2019: Data, Sensing, and Analytics, pp. 209–216. https://doi.org/10.1061/9780784482438.027.
Tan, J.-S., Elbaz, K., Wang, Z.-F., Shen, J.S., Chen, J., 2020. Lessons learnt from bridge collapse: a view of sustainable management. Sustainability 12 (3), 1205. https://doi.org/10.3390/su12031205.
Tiwari, A., Turner, C.J., Majeed, B., 2008. A review of business process mining: state-of-the-art and future trends. Bus. Process Manag. J. https://doi.org/10.1108/14637150810849373.
van der Aalst, W.M., 2010. Process discovery: capturing the invisible. IEEE Comput. Intell. Mag. 5 (1), 28–41. https://doi.org/10.1109/MCI.2009.935307.
Van Der Aalst, W., 2012. Process mining. Commun. ACM 55 (8), 76–83. https://doi.org/10.1145/2229156.2229157.
Van der Aalst, W.M., 2016. Process Mining: Data Science in Action. Springer, 3662498510.
Van der Aalst, W., Weijters, T., Maruster, L., 2004. Workflow mining: discovering process models from event logs. IEEE Trans. Knowl. Data Eng. 16 (9), 1128–1142. https://doi.org/10.1109/TKDE.2004.47.
Van Der Aalst, W.M., Reijers, H.A., Weijters, A.J., van Dongen, B.F., De Medeiros, A.A., Song, M., Verbeek, H., 2007. Business process mining: an industrial application. Inf. Syst. 32 (5), 713–732. https://doi.org/10.1016/j.is.2006.05.003.
Van der Aalst, W., Adriansyah, A., van Dongen, B., 2012. Replaying history on process models for conformance checking and performance analysis. Wiley Interdiscip. Rev. : Data Min. Knowl. Discov. 2 (2), 182–192. https://doi.org/10.1002/widm.1045.
Wang, T.-K., Huang, J., Liao, P.-C., Piao, Y., 2018. Does augmented reality effectively foster visual learning process in construction? An eye-tracking study in steel installation. Adv. Civ. Eng. 2018 https://doi.org/10.1155/2018/2472167.
Wang, Y., Liao, P.-C., Zhang, C., Ren, Y., Sun, X., Tang, P., 2019. Crowdsourced reliable labeling of safety-rule violations on images of complex construction scenes for advanced vision-based workplace safety. Adv. Eng. Inf. 42, 101001 https://doi.org/10.1016/j.aei.2019.101001.
Weijters, A., van Der Aalst, W.M., De Medeiros, A.A., 2006. Process mining with the heuristics miner-algorithm. Technische Universiteit Eindhoven, Tech. Rep. WP 166 (July 2017), 1–34. In: https://research.tue.nl/en/publications/process-mining-with-the-heuristicsminer-algorithm.
Woodcock, K., 2014. Model of safety inspection. Saf. Sci. 62, 145–156. https://doi.org/10.1016/j.ssci.2013.08.021.
Wu, C., Wu, P., Wang, J., Jiang, R., Chen, M., Wang, X., 2021a. Critical review of data-driven decision-making in bridge operation and maintenance. Struct. Infrastruct. Eng. 18 (1), 47–70.
Wu, C., Wu, P., Wang, J., Jiang, R., Chen, M., Wang, X., 2021b. Ontological knowledge base for concrete bridge rehabilitation project management. Autom. Construct. 121, 103428 https://doi.org/10.1016/j.autcon.2020.103428.
Wu, C., Li, X., Guo, Y., Wang, J., Ren, Z., Wang, M., Yang, Z., 2022. Natural language processing for smart construction: current status and future directions. Autom. Construct. 134, 104059.
Wu, C., Li, X., Jiang, R., Guo, Y., Wang, J., Yang, Z., 2023. Graph-based deep learning model for knowledge base completion in constraint management of construction projects. Comput. Aided Civ. Infrastruct. Eng. 38 (6), 702–719. https://doi.org/10.1111/mice.12904.
Xu, Q., Chong, H.-Y., Liao, P.-c., 2019. Exploring eye-tracking searching strategies for construction hazard recognition in a laboratory scene. Saf. Sci. 120, 824–832. https://doi.org/10.1016/j.ssci.2019.08.012.
Zhang, J., Wu, M., Sheng, V.S., 2018. Ensemble learning from crowds. IEEE Trans. Knowl. Data Eng. 31 (8), 1506–1519. https://doi.org/10.1109/TKDE.2018.2860992.
Zheng, Z., Zou, B., Wang, Y., Li, S., Gao, Y., Yang, S., 2020. A temporally-calibrated method for crowdsourcing based mapping of intra-urban PM2. 5 concentrations. J. Clean. Prod. 269, 122347 https://doi.org/10.1016/j.jclepro.2020.122347.