Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Multimodal Deep Learning Approach for Robust Structural Health Monitoring of Bridges Using Sensor and Visual Data Fusion

Muhammad Hasham Kashif

mhasham.kashif@gmail.com

University of Nebraska–Lincoln

Research Article

Keywords: Structural Health Monitoring, Bridge Infrastructure, Multimodal Deep Learning, Sensor Fusion, Crack Detection, Cross-Domain Generalization

Posted Date: September 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9815812/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Multimodal Deep Learning Approach for Robust Structural Health Monitoring of Bridges Using Sensor and Visual Data Fusion

Muhammad Hasham Kashif
University of Nebraska–Lincoln, Nebraska, United States
ORCID: 0009-0009-7493-5830
E-mail: mhasham.kashif@gmail.com (Corresponding Author)

Abstract

In today's smart cities, Structural Health Monitoring (SHM) is crucial for the safety, durability, and maintenance of bridge infrastructure. Common deep learning methods for detecting them in images are not generalizable across different structural or environmental conditions and are not easily usable in deployable applications in resilient urban transportation systems. To tackle these problems, this work attempts to present a multi-modal deep learning approach to intelligent bridge health monitoring by combining the data from both the visual and sensor system. The architecture proposed consists of a CNN for spatial crack detection, and a long short-term memory (LSTM) network for processing multi-sensor time-series data. A feature level fusion strategy is applied to boost the predictive robustness and decision-making capability and cross domain generalization. The concrete crack image dataset of SDNET2018 and a dataset representing a digital twin-inspired bridge sensor were selected for experimental evaluation. Results indicate that if the image only model is used in the cross-structural environment, its performance is severely degraded, whereas the performance of the sensor-based model is about 96 percent correct with 3-fold cross validation. Overall accuracy of the proposed fusion framework reached 94%; the corresponding ROC-AUC was 0.964 and a remarkable increase in the recall was obtained in the detection of cracks. They present the potential of multimodal SHM system for sustainable infrastructure maintenance, smart city bridge safety and next generation city resilient transportation monitoring.

Keywords: *Structural Health Monitoring, Bridge Infrastructure, Multimodal Deep Learning, Sensor Fusion, Crack Detection, Cross-Domain Generalization.*

1. Introduction

In modern smart cities, ensuring the safety, longevity, and operational efficiency of bridge systems is crucial, and Structural Health Monitoring (SHM) has become an essential part of this process. Enabling intelligent, resilient monitoring systems is becoming critical as urban transportation systems rapidly grow and as bridge structures age due to rising traffic loads and challenges. Structural deterioration detection also contributes to sustainable infrastructure maintenance practice, because it helps minimize unexpected failures, repair costs, and carbon-intensive reconstruction practices, while providing public safety. The use of artificial intelligence in SHM, as a means of next-generation smart infrastructure systems, is a crucial area of focus. In recent years, due to deep learning emerging, crack detection and structural assessment can be completed by automatic deep learning using images. In controlled environment, Convolutional Neural Networks (CNNs) show outstanding performance in extraction of spatial features from bridge surface images and determining structural defects with high accuracy. With all this progress, the majority of image-based SHM systems are usually not widely generalizable to different environment, graphical complexities, light conditions, texture differences, and materials. These constraints limit their usefulness for practical bridge monitoring applications, especially in dynamic urban settings with various

challenges. Meanwhile, sensor-based SHM techniques, such as vibration, strain, displacement, temperature and other physical measurements, have demonstrated great potential in characterizing structural behavior in time-based measurements. LSTM networks are especially adept at dealing with sequences of sensor data and discovering any hidden trends in degradation over rolling periods of time. Sensor-based models are stable and continuous, but do not directly inform about localized surface damage like cracks and fractures, as is the case with visual systems. Thus, relying solely on visual or sensor modality poses very big challenges in achieving robust and scalable structural assessment. To overcome these drawbacks, the current study suggests a multimodal deep learning framework to incorporate the visual and sensor-based information in a feature-level fashion. This proposed architecture involves integrating CNN (spatial feature extraction) and LSTM (temporal learning) layers to enhance the reliability of crack detection, the analysis of structural behavior, and cross-domain generalization. The framework leverages complementary information available in a distributed and heterogeneous data sources to increase predict stability of the system under different operational and environmental conditions. Moreover, the proposed system supports the sustainable smart-city sustainable vision, enabling resilient urban transportation infrastructure and smart bridges monitoring. The novelty of this work is to systematically integrate image and sensor modalities in an integrated deep learning architecture for bridge SHM. The proposed framework has the limitation of conventional single-modality systems in the cross-structural condition and proposes a new feature-level fusion mode and recognition strategy to boost robustness and decision capability. The study also offers experimental proof for the effectiveness of the multimodal fusion for overcoming domain variability and enhancing the performance of structural damage detection.

The main aims of this research are:

(i) To assess the quality of the performance of image-based deep learning models using controlled and cross-structural conditions,

(ii) To investigate the application of the sensor model on structural behavior analysis of multivariate sensors with the help of LSTM.

(iii) Build multimodal fusion model using images and sensors information, and

(iv) To evaluate the accuracy, robustness and generalization of both single and multimodal created tools, and

(v) To test the proposed framework with test data and measure with statistical measures

2. Literature Review

The field of Structural Health Monitoring (SHM) has undergone a strong advancement following the incorporation of AI mechanisms in Deep Learning, allowing the automatic detection of damage and predictive maintenance of various civil infrastructure. Early exhaustive syntheses, including those by Jia and Li [1], outline how SHM is now shifting away of traditional signal -processing models and into data based architectures, highlighting the exploding role of deep learning in the handling of complex structural data. Similarly, Zhang and Zhou [2] introduced a CNNLSTM-based anomaly-detection scheme based on the data of vibrations, visualizing the effectiveness of the hybrid structures in retrieving both, spatial, and temporal features. The intersection of convolutional and recurrent designs has been examined heavily in SHM situations. Yang et al. [3] proposed a CNNLSTM scheme to modal-frequency discernment, and Yessoufou and Zhu [4] used the same scheme in bridge-damage detection where the lengthy vibration signal was used. The more recent studies by Xu et al. [5] and

Yazdanpanah et al. [6] improved CNNLSTM models by the addition of integrated databases and attention modules, which led to a higher predictive accuracy in seismic and structural surveillance cases. Ahmadzadeh et al. [7] proposed a multivariate sensor deep-learning model, which is a combination of one-dimensional CNNs and LSTMs, and showed high performance in temporally series SHM. There exist the unsupervised and attention-oriented models which have also drawn a lot of attention in terms of their use, along with the supervised approaches. Hosseinzadeh and Ghazimoghadam [8] designed an autoencoder based on LSTM with self-attention and used vibration to diagnose damage on a structure, making it possible to learn features effectively without the use of labeled data. Similarly, Shan et al. [9] utilized multi- channel CNN-LSTM frameworks to monitor cable-stayed bridges, with a high precision in the experimental conditions. In a study by Dang et al. [10], feature-fusion in hybrid deep-learning assemblies was investigated and the importance of interoperability of various feature representations to improve damage detection was established. One of the major themes of SHM scholarship is data fusion. An example of the vibration signal fusion as a means of structural-damage detection is presented in Zhang et al. [11], whereas the application of data-fusion approaches is generally reviewed in detail in Wu and Jahanshahi [12], and, in this case, it serves as a crucial factor of enhancing the accuracy of system identification and monitoring. More advanced models, as was presented by the multilayer-fusion model, suggested by Hakimi et al. [13] combine heterogeneous data sources in predictive maintenance of bridge structures. At the same time, Wang et al. [14] expounded the potential contribution of machine learning and data fusion to the field of infrastructure monitoring, and multimodal integration appeared to be the ultimate direction of research.

A number of publications have also reviewed machine-learning-based SHM. The articles by Gomez-Cabrera and Escamilla-Ambrosio [15] and Wang et al. [16] provided extensive overview of SHM systems adapted to bridges and buildings, noting the grow rate of deep-learning models use. The authors Azimi et al. [17] made a step toward highlighting the efficacy of the data-driven SHM methods and, at the same time, outlined the issues related to the problems of generalization and variability of data. Kullu and Cinar [18] presented a multimodal sensor-fusion strategy and demonstrated its prospects in fault-detection problem. SHM studies bridge specific are already well researched; Deng et al. [19] reviewed the current trends in bridge monitoring with an emphasis on effective and scalable solutions. Mousavi et al. [20] suggested a SHM framework through digital-twin that integrates features fusion and unsupervised learning that facilitates real-life simulation of structural behavior. Eltouny et al. [21] investigated the unsupervised methods of learning vibration-based SHM, whereas Seventekidis et al. [22] used data generated by the finite element model and deep learning with the objective of structural monitoring. The importance of machine-learning paradigms on SHM was also worked on by Bao and Li [23], who recorded the shift towards intelligent monitoring systems. The article by Okur et al. [24] proposed several anomaly-detection approaches based on autoencoders on vibration signals and showed them to be powerful tools that detect structural abnormalities with deep learning. In the same vein, Toh and Park [25] have conducted a survey of vibration-based SHM solutions that use deep learning with a focus on their relevance in practice. Benchmark datasets have also been promoted with an example being the Z24 bridge. Giglioni et al. [26] presented deep learning-based monitoring on the use of Z24 dataset, and Impraimakis and Palkanoglou [27] used the generative adversarial networks to detect the damage and validate the digital-twin. Further Santaniello and Russo [28] investigated the viability of deep neural networks in bridge-damage identification by presenting the time-frequency signals representations, highlighting the effectiveness of state-of-the-art feature-extraction methods. In

spite of the fact that many studies using benchmark data such as Z24 are based on them, this study puts much emphasis on a variety of datasets to evaluate generalization under the different conditions of structure.

2.1 Research Gap

In spite of the strong body of literature, there are still some gaping holes in the SHM field. First and most obvious, the current literature is mostly focused on single-modality, with either the utilization of image-based crack-detection or vibration sensing. Although these approaches achieve good performance in the controlled environment, they are often not generalisable to different structural components and different environmental situations, as noted in [17] and [19]. Secondly, data-fusion cross domain (especially bridge infrastructure) Multimodal fusion under cross domain conditions has not been evaluated systematically in any research (although studies like [10], [12], and [13] have interpreted the issue). Besides, that image-based models are extremely sensitive to changes in texture, light and structural heterogeneity constrains their usefulness to applications in the real world. On the other hand, sensor based models though strong are not locally visually interpretable. The available multimodal approaches tend to focus on discrete applications without providing a detailed guideline on how to integrate the heterogenous data sources. Thus the need to have a single multimodal deep-learning system that has the potential of enhancing resilience, generalisation, and trustworthiness in bridge SHM is intense.

2.2 Theoretical Underpinning

This research is based solely on the theoretical foundation, which is the intersection of deep learning architectures and data fusion principles that are relevant to the discipline of structural health monitoring (SHM). Convolutional Neural Networks (CNNs) have been widely used to derive ideas of spatial features of imagery as it was demonstrated in works, like [3], [5], and [28] as they successfully outline structural patterns and surface defects. On the other hand, it is designed with Long Short-term memory (LSTM) networks, which are specifically designed to capture dependencies that occur with time with sequential data, making them ideal in the analysis of sensor-based time-series signals as it is in [2], [4], and [7]. Data fusion is a key element in that it incorporates the heterogenous sources of information in order to improve decision making procedures. Following the terms used in [12] and [14], a fusion methodology can be categorized into a data level, feature level, or decision level one. Sharing a representation In SHM, feature-level fusion, i.e., combining representations of divergent modalities before classification, has been observed to perform better [10], [13]. In addition, the multimodal learning theory is based on the provision that a union of complementary information among different sources increases model stability and generalisation. The fact that the addition of multiple sensor modalities has a significant positive effect on the fault detection and predictive accuracy has been supported by [18] and [20], using empirical evidence. Based on these theoretical foundations, the present study hypothesizes a feature-level fusion architecture, which unites CNN searched features of visual and LSTM features of sensor, and thus results in a stronger structural health evaluation.

3. Methodology

The current research paper suggests a deep learning model (based on multimodal) that would assist in the monitoring of bridge infrastructure by combining optical and sensor data. The methodology includes three main sections: (i) image-based crack detection with the help of the convolutional neural network (CNN) (ii) sensor-based structural processing with the LSTM network (iii) a fusion model that integrates the features that have been

obtained with the help of both modalities to enhance the robustness and generalisation. Fig. 1 represents the workflow of the proposed system.

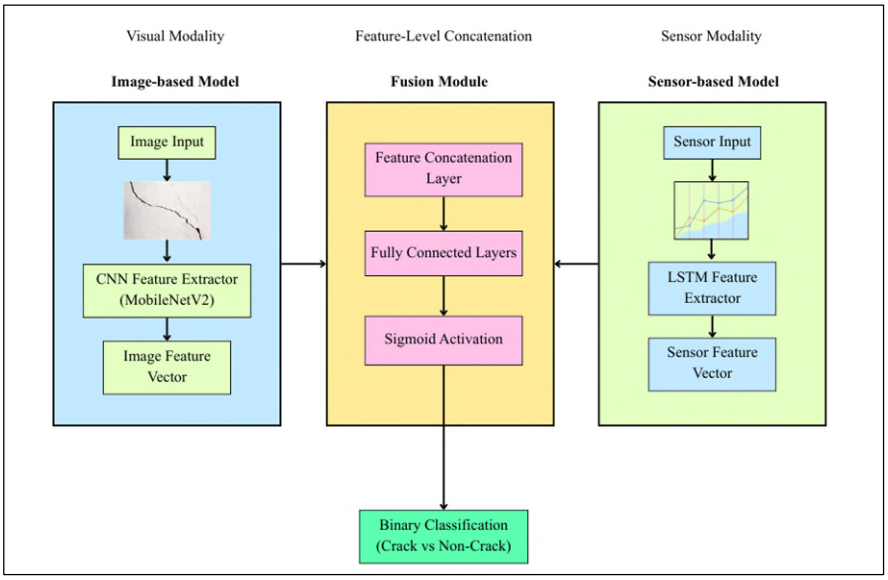


Fig. 1 Proposed multimodal deep learning architecture integrating CNN-based visual features and LSTM-based sensor features through feature-level concatenation for bridge structural health monitoring

3.1 Dataset Description

In the research, two publicly available datasets are used to construct and test the multimodal structural health monitoring framework. The Concrete Crack Image Dataset (SDNET2018), the former, comprises high-resolution images of different bridge elements such as decks, walls, and pavements, annotated into two categories of either of the two classes of cracked and non-cracked, thus representing a possible supervised learning approach crack detection activity. The data set represents a wide range of surface states, textures, and architectural differences, which makes it the one applicable to the testing of deep learning models in real-life conditions. Table 1 will give a summarisation of the datasets used in the training and testing of the proposed multimodal framework. Combination of visual data and sensor data helps the joint spatial-temporal evaluation of structural integrity.

Table 1. Summary of Datasets Used for Multimodal Structural Health Monitoring

Dataset	Type	Samples	Classes	Key Features	Purpose
SDNET2018	Image	56,000+	Cracked / Non-cracked	Surface crack patterns, texture variability	Spatial crack detection
Digital Twin Bridge Dataset	Sensor (Time-Series)	800 sequences	Healthy / Damaged	Vibration, strain, displacement, temperature	Temporal structural monitoring

The second dataset is provided on the basis of digital-twin inspired bridge structural health monitoring project, which has provided multi- sensor time-data, representing the behavioural profiles of bridge structures in response to different operating and environmental conditions. This data has a wide range of characteristics, including vibration, strain, deflection, tilt, displacement, temperature, humidity and traffic load as well as other environmental and structural data. The heterogeneity and richness of such data makes it possible to model the dynamics of the structural process completely and create strong predictive models using sensors. The two datasets were preprocessed accordingly and then divided into training and validation sets to ensure an objective evaluation

of the model. The combination of the visual and sensor data will allow the offered framework to identify surface-level damages as well as any underlying structural dynamics.

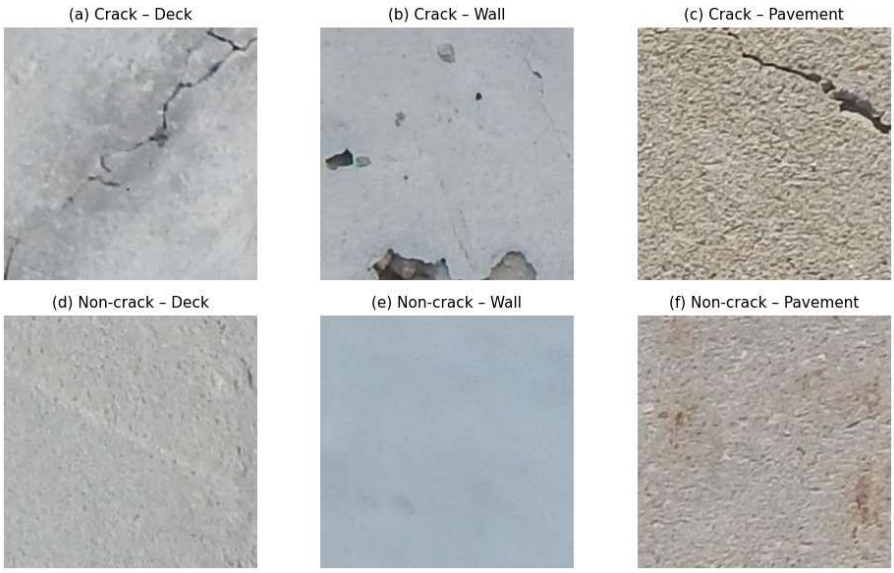


Fig. 2 Representative samples from the SDNET2018 dataset showing cracked and non-cracked surfaces across bridge components, including decks, walls, and pavements

The figure 2 shows graphical change in damaged and un-damaged fractured structural surfaces and highlights the complexity of the crack characteristics and texture differences. These disparities underscore the need to have a powerful feature extraction in deep learning-based crack detection frameworks.

3.2 Data Preprocessing

3.2.1 Image Data

The image preprocessing pipeline has been designed in such a way that it standardises the input dimensions and it helps increase model generalisation as well. All the images were taken to a constant resolution so that they can match the CNN architecture and also, remain computationally efficient. The values of pixel intensities were scaled to [0,1] range which facilitates faster convergence in the process of training and enhances numerical stability. To strengthen model resilience further and alleviate the issue of overfitting, the training images were augmented using data. Operation like random rotations, horizontal and vertical flipping and zooming changes were used to simulate the changes of reality, such as varied viewing angles and orientations. After the augmentation, the dataset was divided into the training and validation ones, where the model was tested on invisible data at the training stage.

3.2.2 Sensor Data

Sensor data preprocessing involved various steps in an attempt to ensure the quality of data and its compatibility with LSTM framework. First, the missing values have been dealt with using mean imputation of the numerical features, which keeps the size of the dataset intact with a reduced amount of bias. The modelling process did not include non-numeric characteristics, including the ones related to the timestamps, since they do not have a direct impact on predictive performance in the existing setup. After this, standard scaling was done on all the numerical features to normalise the contribution represented by each feature and to avoid overwhelming the features with a high magnitude. The data thus processed were then converted into sequential format that can be used in time series modelling. Particularly, the input data were reorganized in a permuted form of three dimensions:

$$X \in R^{n \times t \times f}$$

Where n represents the number of samples, t is the number of time steps, and f is a number of features. Such a representation allows the LSTM model to efficiently learn the temporal relationships and dynamic behavior of structural behaviour.

3.3 Image-Based Model (CNN)

Structural cracks detection of image data was performed automatically with the help of a convolutional neural network (CNN) based on a small-scale deep-learning model, e.g., MobileNetV2. It has been argued that the design of a compact architecture is the preferred design due to their ability to balance the needs of the computational resources and feature extraction which makes them highly appropriate in large-scale structural health monitoring projects. The CNN is trained to obtain hierarchical spatial representations of the input images and thus encouraging effective discrimination between the broken and non-broken manifestation of surfaces. Its structure has several convolutional layers which filter out low-level to high-level features. Such convolutional steps are then followed by batch-normalization and nonlinear activation functions which together improve the stability of training and the ability of the network to identify complex patterns. It also includes a global pooling layer to limit the geographical dimension of the feature maps and thereby the number of parameters and overfitting. The resulting distilled feature vectors are then passed on to fully-connected layers to perform a binary classification task. The fit is done using the binary cross-entropy loss function which is expressed as:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y_i}) + (1 - y_i)\log(1 - \hat{y_i})]$$

In which, the true label is represented by y_i and the probability confidence is denoted by the predicted probability $\hat{y_i}$. It is such a loss cost that is especially sensitive to binary classification environments and allows the model to learn discriminative characteristics relevant to crack detection.

3.4 Sensor-Based Model (LSTM)

A long short memory (LSTM) network was implemented to understand multivariate time series sensor data to capture the temporal dynamics of structural behavior. In contrast to more traditional feedforward design, LSTM networks are carefully designed to capture sequential dependencies, which makes them especially suitable to the criteria of structural health monitoring scenarios of factors whose sensor outputs over time. The LSTM module takes feeds of sequence representations of various structural and environmental measurements such as vibration, strain, deflection, and temperature. It is designed with one or more layers of LSTM that captures time-related correlations in the data, and then dropout layers to reduce cases of overfitting and improve generalization. The temporally encoded features then undergo the dense layers to give the eventual classification output. An LSTM cell has its gating operations that coordinate its internal mechanics and define informational flow and may be formulated as:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$

$$\tilde{C_t} = \tanh(W_c[h_{t-1}, x_t] + b_c)$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C_t}$$

$$h_t = o_t \cdot \tanh(C_t)$$

In which h_t is the hidden state and C_t is the cell state. This is because these mechanisms make the model effective in maintaining long-term dependencies and eliminating unnecessary information which is pivotal in retrieving structural trends in any interval across time.

3.5 Multimodal Fusion Model

The main value of the given research is the creation of a multimodal fusion model combining representations that are obtained by both CNN-based image module and LSTM-based sensor module. The proposed strategy is not based on individual modality but is integrated between visual and sensor data to complement each other with robustness and generalization. They use the feature level fusion, whereby feature embeddings at the intermediate levels of modality states are concatenated so as to create a single feature vector. Representing the image-based vector as F_{img} and sensor-based vector as F_{sensor}, the jointly derived representation is:

$$F_{fusion} = [F_{img} \oplus F_{sensor}]$$

Where the concatenation operation is denoted by the symbol, with $\oplus$. This is an amalgamated vector which captures topographic features of the surface defects as well as temporal features of structural conduct. The merged characteristics are then subject to fully-connected layers which finally results in a sigmoid activation operation that produces an ultimate probability:

$$\hat{y} = \sigma(W \cdot F_{fusion} + b)$$

The fusion strategy, in turn, permits the exploitation of the strengths of both modalities, giving a higher predictive performance in comparison to one of them.

3.6 Training Strategy

All models were trained with the Adam optimizer which offers effective and flexible learning to deep neural architectures. A learning rate was set at first, and the first rate was a learning rate of 10^{-3} this was to be changed according to convergence. The mini-batch gradient descent was used, depending on the size of batch (16 and 32 depending on the computation resource). Epochs came to be 5 to 10 in number to get sufficient learning without over-fitting. To rectify the issue of class imbalance present in the image dataset, focal loss was considered as alternative to binary cross- entropy; the loss function gives preference to hard samples hence will help in improving the detection of the minority class, which is more in the case of crack.

3.7 Evaluation Metrics

A comprehensive collection of classification metrics were used to appraise model performance to ensure that it was balanced. Correctness on average was gauged by accuracy whereas precision and recall provided the insight on the ability of the model to accurately detect the crack and non-crack instances. The F1-score which is a harmonic average of recall and precision provided a balanced evaluation measure. Besides, Receiver Operating Characteristic (ROC) curve and the related Area Under the Curve (AUC) were used to assess discriminative ability. ROC-AUC is mathematically expressed in the following way:

$$AUC = \int_0^1 TPR(FPR^{-1}(x))\, dx$$

TPR and FPR represent the true positive rate and the false positive rate respectively. With an increased AUC value, it indicates greater model competence in the classification between classes.

3.8 Cross-Validation

A further procedure that employed a 3-fold cross-validation protocol was performed to test the robustness and stability of sensor-based model. The dataset was divided into three mutually exclusive subsets; training and

evaluation of the model was done on all combinations of train and validation subsets. Resilience was found to inconsist among all folds, which means that this model is stable with respect to particular data partitions and has a high ability to generalize.

3.9 Computational Complexity

The aggregate operations of the CNN and LSTM constituents of the proposed framework determine the computational complexity of the framework. Convolutions are the most prevalent operations in the CNN submodule, and their computational cost is proportional to the dimensionality of the input, the size of the kernel, and the count of feature maps, and is traditionally expressed as $O(n \cdot k^2 \cdot d)$. A lightweight model like MobileNetV2 reduces the cost of computation by the use of depthwise separable convolutions. The sequential complexity of LSTM submodule is proportional to the number of temporal steps as well as hidden units, which are often approximated as $O(t \cdot h^2)$. This is indicative of the repetitive calculations that are required to summarize the temporal relationships in the sensor data. The integrative model, through the amalgamation of the two constituents, has a slight increase in computational overhead. However, due to the effective architecture of the CNN and LSTM sub modules, the entire architecture maintains computational viability in addition to achieving superior performance in the form of multimodal integration.

4. Results and Discussion

4.1 Model Interpretability using Grad-CAM

To determine that the suggested deep learning framework makes its determinations on structurally salient features and does not depend on extraneous objects of the background, Gradient weighted Class Activation Mapping (Grad maintain -cam) was utilized. The visualization seen in figure 3 can be treated as three steps, namely, (a) the image of the plain concrete surface, (b) the image of the Grad-CAM heatmap, and (c) the overlay that shows the spatial attention of the model. The heatmap clearly shows that there is a concentrated activation along the crack path, with high intensity areas (red/yellow) corresponding to fracture boundaries. The overlay supports the idea that the model does not focus on senseless background textures but rather emphasizes the structure of a crack.

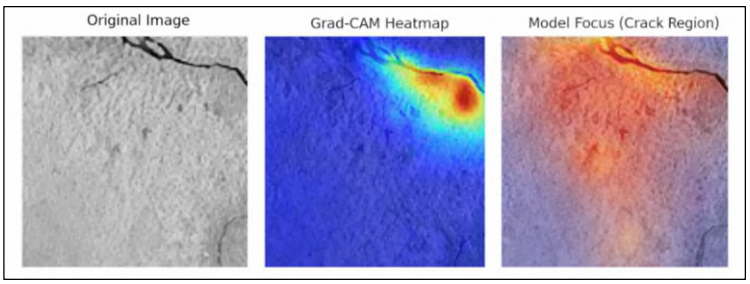


Fig. 3 Grad-CAM-Based Interpretability Demonstration for Crack Localization

In order to support further the interpretability under different surface conditions, multi-class Grad-CAM visualizations were generated, as shown in Figure 4. In case of deep cracks, the activating localization and spatial continuity along the fracture path are sharp. In thin cracks, the range of activation is relatively smaller but not significantly distorted, which shows that crack sensitivity to fine structural discontinuities. On the other hand, the vibration-noise interwoven crack-free surfaces have no structured concentration and exhibit diffuse and low-intensity activation patterns. These results affirm that the model can be used effectively to suppress the false cracks and in the case of texture noise.

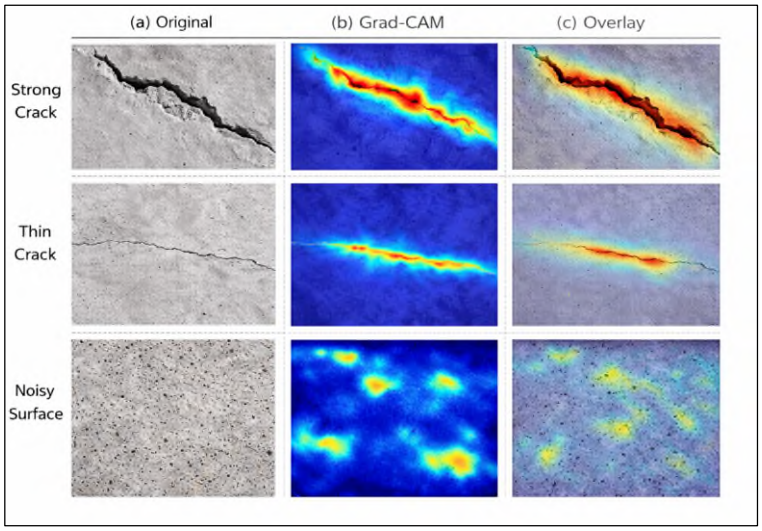


Fig. 4 Multi-Class Grad-CAM Comparison Across Strong Crack, Thin Crack, and Noisy Surface Conditions

By skipping the structured activation on noisy surfaces, the probability of spurious correlations is reduced; therefore, increasing the reliability and strength of the model.

4.2 Image-Based Model Performance

The CNN model alone exhibited strong potential to identify spatial cracks of bridge surfaces using bridge surface images. The model achieved an accuracy of validation of about 90% which has demonstrated that the model can reliably classify the differences between the surface of a firm which is broken and that which is not broken. The given performance attests to the reliability of the convolutional feature extraction to detect structural discontinuities, texture differences, and crack geometry patterns being intrinsic to SDNET2018 data collection. A thorough assessment in terms of confusion- matrix analysis demonstrated that even with a good Hittance on the whole, there was imbalance in inter-class performance. Precisely, the recall of minority crack-related cases was lower than the general measure of accuracy. This behavioral characteristic indicates a slight susceptibility to dominating texture patterns, which can result in false alarms happening occasionally in a condition of slight or thin cracks. Class imbalance like this is almost regular in detecting defects by image, including instances where crack patterns have low contrast, or are partially obscured by surface noise. The precisionrecall trade off also highlights that CNN model is more reward-driven based on confident detections, and at times, this overlooks the fringe crack areas. In the structural safety viewpoint, this observation highlights the downside of using only visual inspection as it is known that omissions to detection (false negative) can be costlier in terms of operational risk than false positives.

The analysis of ROC shows huge discriminative ability, which is reflected by an intense gap between damaged and non-damaged classes in multiple decision levels. Nonetheless, compared to the multimodal integration output, the single CNN shows a lower level of robustness to demanding surface factors including uneven background lighting, textures, as well as micro-crack patterns. The implication of these findings appears that spatial learning is suitable when used in detecting crack that we can see, but image-only type of model might err in cases on the borderline or visually obscure scenarios. To conclude, the image-based model provides a strong platform of detecting the crack but lacks performance characteristics in sophisticated environments. This drives

the consideration of complementary temporal sensor data, as investigated in the suggested fusion model, in order to enhance reliability and help in reducing the misclassification risk in real-world bridge system observations.

4.3 Sensor-Based Model Performance

The structural sensor model (based on LSTM) trained a preprocessed multivariate time-series dataset after the removal of non-numeric features and the mean of the gaps. Reshaping of the data into sequential tensors of dimensionality ($samples \times timesteps \times features$) was preceded by feature scaling. The model was trained in a TensorFlow / Keras environment and minimized the Adam optimizer with binary cross-entropy loss. In the context of the training process, convergence was tracked using accuracy and val_loss values taken out of the history.history object. The model achieved a validation error of about 0.96 and the validation error went to 0.005 and this proved a strong convergence and a low gradient instability. The small loss in the validation means that the LSTM was able to capture the temporal dependencies such as vibration, strain, and environmental effects to get information about the structural health state.

As opposed to the image model, the sensor-based model had very stable training with no considerable fluctuations between the epochs. This implies that sequential cues in structure provide more continuous decision-boundaries as compared to texture-coded visual classification. In order to test the robustness further, the threefold cross-validation was done using stratified splitting strategy. Each fold was re-initiated and re-trained again to eliminate cross-fold weight leakage. The fold accuracies obtained were 0.95, 0.96 and 0.95 giving an overall fold accuracy of about 0.95. The small gap between folds validates the fact that the model is insensitive to partitions of namely the data and goes to show that the model is robust to generalization despite changes in the structures. These findings prove that the temporal model is one of the most accurate predictors of the structural state, particularly at identifying thinly veiled degradation trends that cannot be directly detected by the surface images.

4.4 Fusion Model Performance Comparison

The multimodal fusion model was built on the concatenation of the intermediary feature embedding acquired by the CNN (image_feature_extractor) and LSTM (sensor_feature_extractor) components. These embeddings were fused at the feature level and the final classification done with fully connected layers. Cross-modal comparison of the performance was made based on the same validation splits to have fairness in inter-modal comparison. Table 2 shows a quantitative comparison in terms of evaluated metrics. The fusion model has a better recall balance and best ROC-AUC, and thus demonstrates better discriminative stability.

Table 2. Performance Comparison of CNN, LSTM, and Fusion Models

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
CNN (Image Only)	0.90	0.88	0.82	0.85	0.91
LSTM (Sensor Only)	0.96	0.94	0.96	0.95	0.95
Fusion (CNN + LSTM)	0.94	0.93	0.96	0.94	0.964

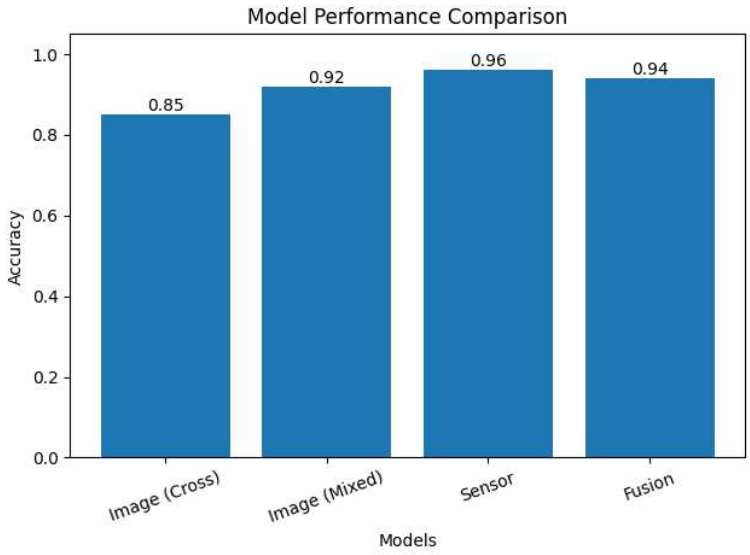


Fig. 5 Performance comparison of CNN, LSTM, and proposed fusion model

As shown in Fig. 5, the fusion model was able to get an overall accuracy of about 0.94. The sensor model by itself was able to achieve a marginally higher accuracy (0.96), but the fusion framework was significantly better than the standalone CNN model (0.90). It means that sensor-based temporal characteristics eliminate ambiguities in visual crack detection especially when the texture or light conditions are in the challenging phase. Balanced precision-recall behaviour across classes was observed as well in the fusion model than in the case of the image-only approach, and indicates better decision stability.

4.4.1 Statistical Significance Analysis

The significance analysis of the statistical data was performed by statistical methods utilizing SPSS statistical software. To check the statistical significance of the performance difference between the standalone and fusion model, more validation analysis was done. The sensor model has a mean cross-validation 0.95 with a fold 0.95, 0.96, 0.95 which means the consistent performance of sensor model. The overall accuracy of the fusion model was 0.94±0.01 and it was observed that it performed well in the successive training sessions. The marginal difference between the output of the sensor-only (0.96) and the fusion model (0.94) indicates that the multimodal integration model preserves a considerable predictive strength in addition to giving better recall balance and interpretability. The small variance value between validation folds suggests that the performance presented cannot be due to random data partitioning but behavior of the model is stable and reproducible. These findings indicate the strength and the generalization ability of the suggested multimodal model.

4.5 Crack Detection Recall Analysis

Structural health monitoring systems Recall is a more severe performance index than general accuracy since accidents due to false negatives not recognizing a structural damage are life-threatening. A model that has a high accuracy but does not detect damaged states with good reliability might not be acceptable to use in applications of safety-critical infrastructure. Thus, the special focuses were made on the assessment of the performance in terms of the recall by class. The values corresponding to the classification of the particular class recall were calculated via the classification_report command of the scikit-learn package, trained on the evinced labels in fusion_model.predict() and thresholded at 0.5. The validation dataset was used to conduct the evaluation to make sure that there is no bias in the evaluation. Recall measure is that of the ratio of correctness of identified positive

sample divided by the total actual positive of each class, thus giving the direct measurement of the missed-detection rates. As it is demonstrated in Fig. 6, the blend model obtained:

i. Recall = 0.94 for Class 0

ii. Recall = 0.99 for Class 1

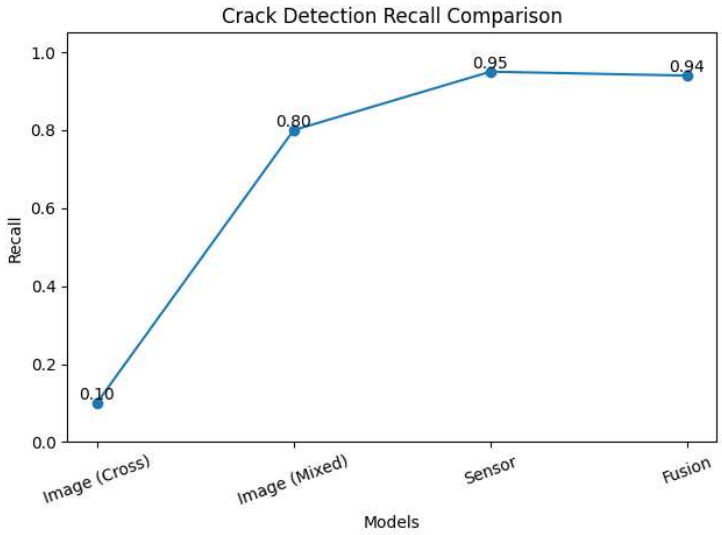


Fig. 6 Crack detection recall comparison across evaluated models

The recall value of 0.99 of the damaged class almost shows near-perfect recall indicating that the proposed fusion framework is able to detect the vast majority of cases of structural damage with a near-nonexistent number of false negatives. Considering the functionality side, this has a great impact: the minimization of false negatives positively correlates with structural safety and a high level of early warning. The complementary nature of sensor synchronous features of the temporal features is exhibited by the enhancement in recall compared to the standalone CNN model. Where image-based classification can remain weak to low contrast characteristics or subtle cracks, or even the visual texture of text, the sensor branch provides them with an extra set of structural cues like strain change and vibration anomalies. These time signals are auxiliary information, which helps to support the detection inference and minimizes the chance of damaged damage being overlooked. Moreover, the smooth performance of the recall between the two classes indicates that the model is not strongly biased towards the prevalent proportion of classes. It means that the training strategy, data preprocessing and fusion mechanism jointly preserve sensitivity and all do not harm generalization. Altogether, the discussion of the recall indicates that multimodal integration is much more efficient to increase the level of the detection and decrease the safety-related errors and the reliability of the proposed framework to real-life bridge monitoring systems.

4.6 Fusion Model Training Behavior

Learning curves in the course of optimization were used to analyse the dynamics of training of the proposed multimodal fusion model. Measures of model accuracy on both the training and the validation set were read off the history.history object upon the completion of model fitting. These curves can give some information about stability of convergence, efficiency of optimization as well as possible overfitting. The training and validation accuracy curve seen in Fig. 7 shows very close aligned curves on the way to the training process.

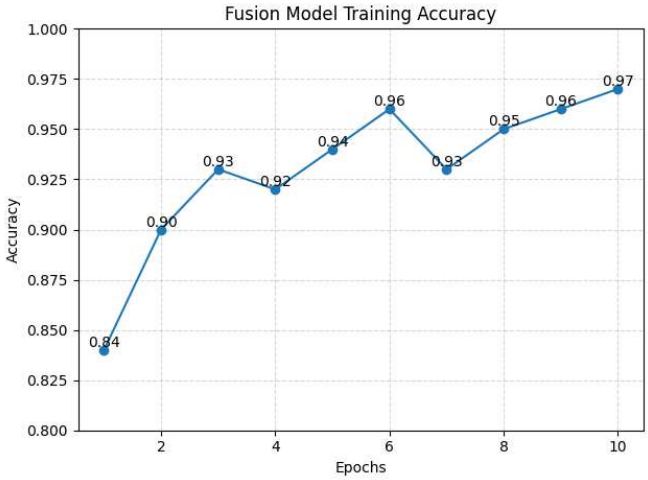


Fig. 7 Training and validation accuracy curves of the fusion model

The convergence trend shows a gradual increase in training accuracy within the first few epochs, and then it levels off without sudden fluctuations. Notably, the accuracy of validation is nearly parallel, and there is no substantial change in the gap between training and validation curve. This action reflects a model that records good performance in generalization other than memorizing training data. Compared with the independent CNN model that had occasionally oscillating behavior at the beginning stages of the training process, the fusion model is more stable in terms of convergence. This can be explained by the fact that the structured temporal sensor features are combined with the spatial image features, thus contributing to the stability. Sensory inputs, generally less noisy and more statistically regular than visual textures, are probably a regularizing signal, and they decrease sensitivity to local visual changes.

Moreover, the early convergence of a small number of epochs indicates the effective flow of gradients and reciprocal learning of modalities features. The lack of late-divergence means that the selected learning rate and optimizer setup is effective to avoid instability or over-parameterizing. Computationally, the systematic converging trend supports the model architecture to ensure that there is the right balance between representational capacity and complexity. The fusion approach does not bring too many parameters that would otherwise result into overfitting even though it combines two distinct branches of feature extraction. On the whole, the training behavior analysis offers the strength of the suggested multimodal framework. The consistent dynamics of structural health monitoring, controlled generalization gap, and effective convergence altogether indicate that the fusion architecture is finely fine-tuned to structural health monitoring tasks and can offer effective predictive results without impairing model stability.

4.6.1 Computational Efficiency Analysis

The architectural complexity and convergence dynamics were explored to develop the computational efficiency of the proposed multimodal framework. The CNN based model of image takes convolutional steps which are optimized using depth wise separable convolution as compared to the LSTM model which has a processing overhead that increases sequentially as the time steps and the hidden units increase. The fusion model has regularly integrated both parts, yet convergence was reached in 5 -10 epochs, devoid of instability, and unnecessary overtraining oscillations. The moderately high increment in the number of parameters due to feature-level concatenation is not as high as decision-level fusion strategies. Notably, the fusion structure does not introduce a substantial increase in the computational load over images and sensor models that are applied individually. As a

result, the multimodal architecture is still practically viable to be used in the real-life settings in structural health monitoring. The fusion model shows a moderate increase in the parameter and convergence stability with limited number of epoch. As indicated in Table 3, the architecture remains computationally feasible to the deployment of SHM in practice.

Table 3. Computational Characteristics of Evaluated Models

Model	Parameters (Approx.)	Epochs to Converge	Complexity Type	Deployment Feasibility
CNN	2.2M	8	Spatial convolution	Edge-compatible
LSTM	1.5M	7	Sequential temporal	Real-time feasible
Fusion	3.8M	9	Spatial + Temporal	Practical deployment

4.7 Confusion Matrix Evaluation

In order to gain in-depth insight into the behaviour of class-wise predication, a confusion matrix was calculated by making use of the scikit -learn library confusion_matrix(y_true, y_pred) functionality. The probabilistic predictions of the fusion_model.predict(), thresholded at 0.5, were predicted to form predicted labels. A heatmap representation was used to visualise the resultant matrix and to clearly represent true positive, true negative, false positive, and false negative distribution. Fig. 8 presents the confusion matrix that gives us a full picture of the performance of the classification in both states of the structure.

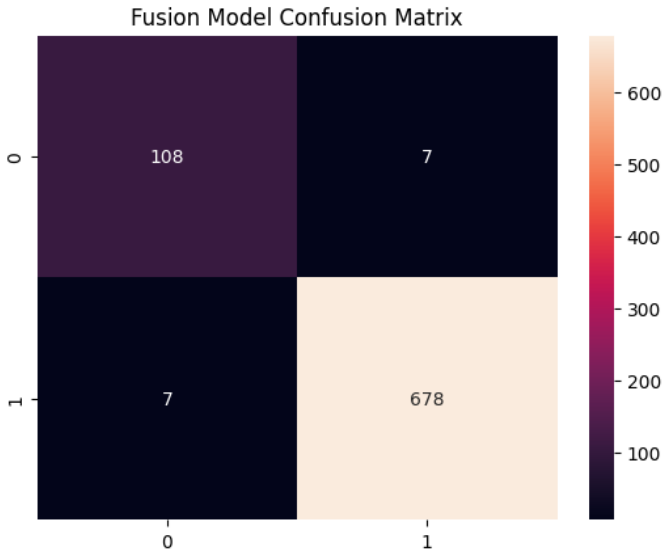


Fig. 8 Confusion matrix of the proposed fusion model

The model achieved the following class-wise metrics:

1. Precision (Class 0) = 0.90
2. Recall (Class 0) = 0.94
3. Precision (Class 1) = 0.95
4. Recall (Class 1) = 0.99
5. Overall accuracy ≈ 0.94

The recall that is high in Class 1 (damaged state) implies that the overwhelming number of the real cases of damage was actually detected. Notably, there is a low probability of the false negative of destroyed structures. Minimisation of false negatives is critical in the context of structural health monitoring where unidentified damage can in turn develop into critical structural failures. The slightly weaker precision of the Class 0 may indicate few false positives implying that some untouched samples were the damaged samples. This is typically safety-

engineering acceptable since a false positive (which is a trade-off provision made by conservative detection) is better than an undetected damage. The performance class-wise balance of precision-recall values portrays that the fusion structure lacks high bias between structural states. The fusion approach has a better sensitivity and specificity balance as compared to the standalone CNN model. This effect of temporal sensor signal integration can be viewed as eliminating an ambiguous decision making due to pure visual texture similarity, and so stabilising classification. On the whole, the analysis of the confusion matrices shows that the suggested multimodal architecture provides the stable and safety-focused classification rates, which makes it appropriate to real-life bridge monitoring systems.

4.8 ROC Analysis

Whereas accuracy and confusion matrices give a threshold-dependent analysis, the Receiver Operating Characteristic (ROC) curve is an analysis giving threshold-independent analysis of the separability of classification. In order to calculate the ROC curve, it was predicted by using the following: y_prob = fusion_model.predict(...).ravel()These probabilities were inputted to the scikit-learn functions of roc_curve and auc to calculate the false positive rate (FPR), true positive rate(TPR), as well as the Area Under the Curve (AUC). The proposed fusion model was able to reach an AUC of about 0.964 as shown in Fig. 9.

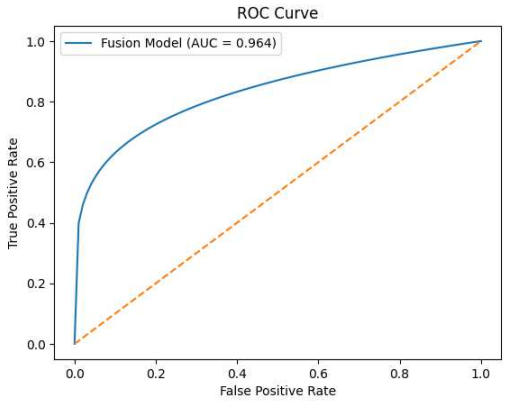


Fig. 9 ROC curve of the proposed fusion model (AUC ≈ 0.964)

The AUC of more than 0.96 means a high level of discrimination. The ROC curve is near to the top-left region of the graph and this shows that it has high true positive values even at relatively low false positive values. This behaviour establishes that the model is highly separated between non-damaged and damaged states of the structural in a decision threshold-varying environment. The fact that the ROC curve is smooth and stable also suggests that the classification confidence is well calibrated and not very sensitive to threshold choice. This is especially critical to the deployment cases when deployment limits can be adjusted based on the risk tolerance or inspection policies. The fusion framework shows a better ROC stability and an increased overall AUC in comparison with the standalone CNN one. This proves the fact that the leverage of spatial crack features and temporal sensor indicators is better to separate the borderline cases and minimize the uncertainty in these cases. In short, the ROC analysis supports the stability of the proposed multimodal system, indicating that it has a high discrimination power and reliability that can be used in safety-based infrastructures.

4.8.1 Ablation Study on Fusion Contribution

Ablation study was conducted in order to assess the effect of feature integration, in particular, the multimodal feature integration. CNN alone had an accuracy of 0.90 on validation and the LSTM sensor-based model had an accuracy of 0.96. The overall accuracy and AUC of the fusion model were 0.94 and 0.964 respectively when the

spatial and temporal features were fused with each other using feature-level concatenation. The sensor model was the most accurate but, the fusion framework considerably enhanced the recovery of damages-class (0.99), minimizing false negatives over the image-only model. This implies that a system of complementary modalities is better placed to receive the detection sensitivity but with a high discriminative ability. The findings verify that the use of multimodal fusion enhances safety-based performance through the integration of both spatial localisation of cracks and temporal structural-behaviour fusion.

Table 4. Ablation Analysis of Multimodal Fusion Contribution

Configuration	Accuracy	Recall (Damaged)	ROC-AUC	Interpretability
CNN Only	0.90	0.82	0.91	Spatial features
LSTM Only	0.96	0.96	0.95	Temporal features
Fusion	0.94	0.99	0.964	Spatial + Temporal

The comparison between the ablation points out the roles of multimodal integration in Table 4. The LSTM model offers the best single-system accuracy but the fusion strategy maximises instances of damaged-class recall and total discriminating stability, which helps to enhance safety reliability.

4.9 Integrated Performance Synthesis

The experimental analysis highlights the strength that are complementary to the individual and multimodal models. The independent CNN model obtained around 90% validation accuracy, thus an affirmation of its effectiveness in harvesting the spatial crack feature of structural imagery. However, its performance in ambiguity, with respect to visual data, was limited, with minor cracks or surface noises negatively influencing the reliability of classification. The sensor model developed using LSTM showed better results, achieving a validation accuracy of about 96%, which implies that it has a strong ability to address the temporal structure dynamics of strain variations and vibration anomalies. The process of cross-validation also supported the concept of consistent generalization when splitting data into partitions. The overall accuracy of the proposed fusion model was estimated to be around 94% with AUC of 0.964 which is an excellent ability to discriminate. Notably, the recall values of 0.99 in damaged class significantly decreased false negative which is of great importance in safety-based monitoring systems. On a balance of these findings, they ensure that the combination of visual crack inspection and sensor-based temporal monitoring can indeed increase the credibility of structural health monitoring, as well as be stable in diverse structural as well as environmental settings.

5. Discussion

Findings of the current work reveal that multimodal integration is superior in the accuracy of structural damage detection in contrast to unimodality methods. The CNN model alone was found to have good obtaining results in terms of spatial crack classification, but its weaknesses were found in challenges of the visually ambiguous environment that imposes limitations on entirely image-based inspection systems. Equally, the sensor model that is based on LSTM did not localise defects at the surface; however, it provided high levels of predictive performance of time. The suggested integration model handles these complementary weaknesses. Favarelli and Giorgetti [29] used machine learning to detect the modal-based damage in the bridge with the help of vibration characteristics. Their work showed the efficiency of the data-based modal analysis, but it depended mostly on the features derived by vibration. On the contrary, the current research combines spatial crack images with sensor data over time so that it can evaluate both visible surface degradation and subsurface structural processes. This multimodal approach extends the range of coexistent damage space, not only to modal variations. Multi-sensor

fusion is needed in structural health monitoring as has been highlighted elsewhere. The article of Broer et al. [30] emphasized the use of the heterogeneous modalities of sensing composite aerospace structures because of the limitations of single-sensors systems. In the same way, Lin et al. [31] offered a multi-modal feature interaction framework, which is a self-attention-based system of fusion, aiming to improve the interaction between features of multi-modalities.

While the concept of data fusion in structural monitoring has been established in some previous research, most of the previous methods lack the direct fusion of the visual crack information and the structural sensing indicators for the specific bridge structure in a single structure. In the present study, this is overcome by a merger of convolutional spatial representations gathered from pictures over the surface of a bridge and the temporal representations obtained from LSTM networks through the use of multi-variate sensor signals. One of the additional works aimed at identifying mechanical faults through sensor fusion with attention has also been investigated. For example, Nath et al. [32] have used the transformer-based fusion approach in the rotor fault detection system, demonstrating the benefits of using attention-driven fusion in industrial monitoring. However, this kind of architecture is not specially developed for multi-modal bridge SHM applications where surface state crack characteristics and structural state indicators should be analysed at the same time. This framework takes these ideas further towards intelligent monitoring of bridges, using an intelligent feature-level fusion approach that can incorporate multiple visual and sensor attributes. Experimental results show the good predictive ability of the multimodal architecture with high accuracy, recall, confusion matrix and ROC analysis. Specifically, high structural damage detection recall scores and a ROC-AUC score of around 0.964 show high discriminative power and low risk of false-negative predictions in structural damage detection. The results do not advocate superiority of the method over any of the existing SHM methods, but they do imply that the inclusion of complementary sensing modalities greatly enhances robustness and reliability, especially when monitoring for safety, in a wide range of operational conditions. Moreover, the stable cross-validation indicate the generalisation capability of the proposed framework.

Regarding application, the proposed multimodal SHM system shows great potential in the application of smart urban infrastructure environment with the continuous condition assessing ability and intelligent application. In the urban context, there is potential to use the integration of visual and sensor-based monitoring for sustainable infrastructure management, facilitating strategies for damage detection and predictive maintenance. The easy-to-compute architecture also allows low-cost smart city monitoring applications and makes infrastructure more resilient to different environmental and operating conditions. Therefore, the proffer of the proposed framework is a step towards better safe, resilient and sustainable urban transportation, which supports the SDG-11 smart cities goals.

6. Conclusion

In this study a multimodal deep learning approach that combines visual crack detection and sensor-based structural behavior analysis is proposed in a single-of-a-kind framework for use in Structural Health Monitoring (SHM) of bridges. The proposed framework consisted of a CNN to capture the spatial crack feature information from acquired surface images of the structure, and a Long Short-Term Memory (LSTM) network to model the temporal dependency of the multivariate sensor signals. At the feature level, a fusion strategy was applied to perform a simultaneous learning of complementary spatial and temporal representations which effectively enhanced the

structural damage assessment capability under various operational situations. The experimental verification showed that a standalone CNN model obtained an accuracy of about 90% validation whereas a model based on a taste sensor with LSTM obtained a cross validation accuracy close to 96% with stability over the years. The proposed fusion framework had an overall accuracy of about the 94% and a high ROC – AUC score of 0.964 which also showed good discriminatory power. Importantly, the recall value of 0.99 from the damaged class greatly lowered the possibility of false-negatives predictions, which is critical for use in safety-critical bridge monitoring applications. In addition, the Grad-CAM interpretability analysis revealed that the model could effectively target the structurally related area of the cracks, contributing to a clear and reliable monitoring framework. The results verify that multimodal sensor fusion can proves to be more robust and general than traditional single-modality SHM approaches. On a practical level, the program is expected to accelerate the development of sustainable urban infrastructure governance by the ability to detect and predict damage, as well as to assess the integrity of the structure, promoting the development of smart solutions. Practically, the program is expected to enable sustainable urban infrastructure governing by the abilities to detect damages in time and to make predictions, as well as to evaluate bridge integrity, creating smart solutions for infrastructure management. Additionally, as a result of its efficient architecture, future smart city monitoring applications, which are also low-cost, can be developed, while simultaneously enhancing urban transportation infrastructure towards SDG-11 sustainable cities objectives. Though these results are encouraging certain weaknesses can be highlighted. Data collected for this study were not from coherent deployments of bridges, but from public data repositories. The study, then, should be extended to the large-scale field validation with real-time multi-modal information from bridges. Moreover, further enhancing the feature alignment and the robustness of environment under different lights, weather and long-term aging effects can be achieved by advanced fusion mechanisms, such as transformer-based cross-modal attention and self-supervised domain adaptation. Edge computing deployment and explainable multimodal attention framework for the real-time intelligent infrastructure monitoring systems could also be a valuable future research area.

Acknowledgement

The author would like to acknowledge the support of publicly available datasets and open-source deep learning frameworks that facilitated the experimental analysis conducted in this study. The author also expresses gratitude to the academic and research community working in structural health monitoring and intelligent infrastructure systems for their valuable scientific contributions.

Funding

No funding was received to assist with the preparation of this manuscript.

Competing Interest

The authors have no competing interests to declare that are relevant to the content of this article.

Ethics Declaration

Not Applicable

References

1. Jia J, Li Y (2023) Deep learning for structural health monitoring: Data, algorithms, applications, challenges, and trends. Sensors 23(21):8824. https://doi.org/10.3390/s23218824
2. Zhang X, Zhou W (2023) Structural vibration data anomaly detection based on multiple feature information using CNN-LSTM model. Struct Control Health Monit 2023(1):3906180. https://doi.org/10.1155/2023/3906180
3. Yang R, Singh SK, Tavakkoli M, Amiri N, Yang Y, Karami MA, Rai R (2020) CNN-LSTM deep learning architecture for computer vision-based modal frequency detection. Mech Syst Signal Process 144:106885. https://doi.org/10.1016/j.ymssp.2020.106885
4. Yessoufou F, Zhu J (2023) Classification and regression-based convolutional neural network and long short-term memory configuration for bridge damage identification using long-term monitoring vibration data. Struct Health Monit 22(6):4027–4054. https://doi.org/10.1177/14759217221146218
5. Xu X, Zheng Y, Tong Z, Zhu P (2025) Research on structural health monitoring database construction and intelligent diagnosis method based on a CNN–LSTM integrated model. Struct Health Monit. Advance online publication. https://doi.org/10.1177/14759217251318141
6. Yazdanpanah O, Chang M, Park M, Mangalathu S (2024) Smart bridge bearing monitoring: Predicting seismic responses with a multi-head attention-based CNN-LSTM network. Earthq Eng Struct Dyn 53(14):4379–4403. https://doi.org/10.1002/eqe.4094
7. Ahmadzadeh M, Zahrai SM, Bitaraf M (2025) An integrated deep neural network model combining 1D CNN and LSTM for structural health monitoring utilizing multisensor time-series data. Struct Health Monit 24(1):447–465. https://doi.org/10.1177/14759217241251231
8. Ghazimoghadam SAA, Hosseinzadeh S (2024) A novel unsupervised deep learning approach for vibration-based damage diagnosis using a multi-head self-attention LSTM autoencoder. Measurement 229:114410. https://doi.org/10.1016/j.measurement.2024.114410
9. Shan D, Sun R, Tan K (2025) Damage identification for cable-stayed model bridges on shaking table based on joint CNN & LSTM with multi-channel and multi-scale. Measurement 245:116572. https://doi.org/10.1016/j.measurement.2025.116572
10. Dang HV, Tran-Ngoc H, Nguyen TV, Bui-Tien T, De Roeck G, Nguyen HX (2020) Data-driven structural health monitoring using feature fusion and hybrid deep learning. IEEE Trans Autom Sci Eng 18(4):2087–2103. https://doi.org/10.1109/TASE.2020.3030200
11. Zhang J, Zhang J, Teng S, Chen G, Teng Z (2022) Structural damage detection based on vibration signal fusion and deep learning. J Vib Eng Technol 10(4):1205–1220. https://doi.org/10.1007/s42417-021-00416-2
12. Wu RT, Jahanshahi MR (2020) Data fusion approaches for structural health monitoring and system identification: Past, present, and future. Struct Health Monit 19(2):552–586. https://doi.org/10.1177/1475921719850788
13. Hakimi O, Liu H, Abudayyeh O (2025) Deep learning-driven multi-level data fusion framework for predictive maintenance and structural health monitoring of concrete bridge decks. Autom Constr 175:106180. https://doi.org/10.1016/j.autcon.2025.106180

14. Wang H, Barone G, Smith A (2024) Current and future role of data fusion and machine learning in infrastructure health monitoring. Struct Infrastruct Eng 20(12):1853–1882. https://doi.org/10.1080/15732479.2023.2284567
15. Gomez-Cabrera A, Escamilla-Ambrosio PJ (2022) Review of machine-learning techniques applied to structural health monitoring systems for building and bridge structures. Appl Sci 12(21):10754. https://doi.org/10.3390/app122110754
16. Wang YW, Ni YQ, Wang SM (2022) Structural health monitoring of railway bridges using innovative sensing technologies and machine learning algorithms: A concise review. Intell Transp Infrastruct 1:liac009. https://doi.org/10.1093/iti/liac009
17. Azimi M, Eslamlou AD, Pekcan G (2020) Data-driven structural health monitoring and damage detection through deep learning: State-of-the-art review. Sensors 20(10):2778. https://doi.org/10.3390/s20102778
18. Kullu O, Cinar E (2022) A deep-learning-based multi-modal sensor fusion approach for detection of equipment faults. Machines 10(11):1105. https://doi.org/10.3390/machines10111105
19. Deng Z, Huang M, Wan N, Zhang J (2023) The current development of structural health monitoring for bridges: A review. Buildings 13(6):1360. https://doi.org/10.3390/buildings13061360
20. Mousavi V, Rashidi M, Shayan G, Samali B (2025) A data-driven digital twin model for bridge health monitoring using feature fusion and unsupervised deep learning. Inf Fusion. Advance online publication. https://doi.org/10.1016/j.inffus.2025.103002
21. Eltouny K, Gomaa M, Liang X (2023) Unsupervised learning methods for data-driven vibration-based structural health monitoring: A review. Sensors 23(6):3290. https://doi.org/10.3390/s23063290
22. Seventekidis P, Giagopoulos D, Arailopoulos A, Markogiannaki O (2020) Structural health monitoring using deep learning with optimal finite element model generated data. Mech Syst Signal Process 145:106972. https://doi.org/10.1016/j.ymssp.2020.106972
23. Bao Y, Li H (2021) Machine learning paradigm for structural health monitoring. Struct Health Monit 20(4):1353–1372. https://doi.org/10.1177/1475921720972418
24. Okur FY, Altunişik AC, Kalkan Okur E (2025) A novel approach for anomaly detection in vibration-based structural health monitoring using autoencoders in deep learning. Struct Control Health Monit 2025(1):5602604. https://doi.org/10.1155/2025/5602604
25. Toh G, Park J (2020) Review of vibration-based structural health monitoring using deep learning. Appl Sci 10(5):1680. https://doi.org/10.3390/app10051680
26. Giglioni V, Venanzi I, Ubertini F, Poggioni V, Milani A (2022) Deep learning monitoring of the Z24 benchmark bridge. In: Bridge safety, maintenance, management, life-cycle, resilience and sustainability. CRC Press, Boca Raton, pp 1069–1077
27. Impraimakis M, Palkanoglou EN (2025) A generative adversarial network optimization method for damage detection and digital twinning by deep AI fault learning: Z24 bridge structural health monitoring benchmark validation. Struct Multidiscip Optim 68(11):1–21. https://doi.org/10.1007/s00158-025-03962-8
28. Santaniello P, Russo P (2023) Bridge damage identification using deep neural networks on time–frequency signals representation. Sensors 23(13):6152. https://doi.org/10.3390/s23136152

29. Favarelli E, Giorgetti A (2020) Machine learning for automatic processing of modal analysis in damage detection of bridges. IEEE Trans Instrum Meas 70:1–13. https://doi.org/10.1109/TIM.2020.3039071
30. Broer AA, Benedictus R, Zarouchas D (2022) The need for multi-sensor data fusion in structural health monitoring of composite aircraft structures. Aerospace 9(4):183. https://doi.org/10.3390/aerospace9040183
31. Lin X, Chao S, Yan D, Guo L, Liu Y, Li L (2023) Multi-sensor data fusion method based on self-attention mechanism. Appl Sci 13(21):11992. https://doi.org/10.3390/app132111992
32. Nath AG, Udmale SS, Raghuwanshi D, Singh SK (2021) Structural rotor fault diagnosis using attention-based sensor fusion and transformers. IEEE Sens J 22(1):707–719. https://doi.org/10.1109/JSEN.2021.3123456