Received: 9 November 2020 | Revised: 9 September 2021 | Accepted: 5 November 2021

DOI: 10.1002/stc.2892

RESEARCH ARTICLE

WILEY

Continuous video stream pixel sensor: A CNN-LSTM based deep learning approach for mode shape prediction

Ruoyu Yang[1] | Shubhendu Kumar Singh[1] | Mostafa Tavakkoli[2] | Nikta Amiri[2] | M. Amin Karami[2] | Rahul Rai[1]

[1]Department of Automotive Engineering, Clemson University, Greenville, South Carolina, USA

[2]Department of Mechanical and Aerospace Engineering, University at Buffalo, Buffalo, New York, USA

Correspondence
Rahul Rai, Department of Automotive Engineering, Clemson University, 4 Research Drive, Greenville, SC 29607, USA.
Email: rrai@clemson.edu

Funding information
Defense Advanced Research Projects Agency, Grant/Award Number: HR00111890037

Abstract

Modal analysis has emerged as a globally accepted tool to formulate and optimize the behavioral functions of engineering structures, which assists in assessing structural failure and laying out a plan for their maintenance. Modal analysis aims at determining the frequencies, damping ratios, and mode shapes of the system under excitation. However, conventional mode shape measurement methods like contact sensors are prone to precision and accuracy issues owing to the sensor's weight and low spatial resolution. In this paper, we improve upon various existing methods for mode shape determination and introduce the idea of a full-field pixel sensor for mode shape prediction. The proposed computer vision-based deep learning architecture predicts the mode shape of a vibrating structure with significant precision. Besides, a ModeShape dataset consisting of the vibration recording video and finite element analysis (FEA) based label has been curated. Specifically, we introduce a convolutional neural network, long short-term memory (CNN-LSTM) computer vision-based non-contact vibration measurement technique for automated mode shape prediction. The key idea is to use each pixel of a RGB camera as a sensor and process the captured spatio-temporal data to enable mode shape prediction. Our CNN-LSTM model takes the video streams of a vibrating structure as input and yields the fundamental mode shapes. The proposed technique is non-invasive and can extract information at relatively high spatial density. The CNN-LSTM model is proficient by utilizing experimental outcomes. The robustness of the deep learning model has been scrutinized by utilizing specimens of an assortment of different materials and fluctuating dimensions.

KEYWORDS

CNN (convolutional neural networks), computer vision, LSTM (long short-term memory networks), mode shape, modal analysis

Ruoyu Yang and Shubhendu Kumar Singh contributed equally to this work.

https://doi.org/10.1002/stc.2892

© 2021 John Wiley & Sons, Ltd.

1 | INTRODUCTION

Modal analysis (MA) is the study of the dynamic properties of systems in the frequency domain. In recent decades, MA has emerged as one of the prominent tools to optimize and refine the dynamic characteristics of vibrating engineering structures. Structures, including bridges, buildings, dams, pipelines, aircraft, ships, among others, are complex engineered systems forming an integral part of our society and ensuring our economic and social well-being.[1] To evaluate the remaining useful life (RUL) and monitor damage occurrence in a structure, subject to the severe working conditions, various government bodies, like the department of transportation and construction, use structural health monitoring (SHM) methods to ensure public safety.[2] The modal analysis finds extensive application in vibration-based structural health monitoring (SHM) of bridges[3] and wind turbines.[4] Apart from structural engineering, MA has also found increasing applications in mechanical engineering, aeronautics, acoustics, space structures, and bio-mechanical engineering. One of the essential components of MA is the modal shape determination. Mode shape is a dynamic structural property, which is defined as a specific pattern of vibrations underwent by particles, of a mechanical system, at a specific frequency, and directly reflects the proportion of structural damage. Fundamentally, damage leads to a change in the structure and subsequently brings about a change in structural properties, such as stiffness and mass, and in turn, affects the mode shape.[5] Thus mode shape forms one of the most critical dynamic features of the structure, incorporating damage information that helps in diagnostics and prognostics of structural health.

The vibration-based SHM falls into four broad categories: natural frequency-based method, curvature mode shape-based method, mode shape-based method, and the methods using both mode shapes and frequencies.[6] Each of these techniques has its own sets of disadvantages. There are two main limitations of the frequency-based SHM method. First, at times, significant damage may give rise to insignificant changes in natural frequencies, particularly for large structures, which makes it difficult to detect any defect. Second, measurement errors or variations in ambient conditions can cause uncertainty in measured frequencies.[7] The curvature-based method is more sensitive to the small defects and can solve the problems associated with the frequency-based SHM method. However, for the higher modes, the difference in modal curvature generates several peaks not only at the damage location but also at other positions, which may lead to a false signal of damage. Besides, if the dataset is collected considering a single mode, then there may be fallacious damage indications.[8] Compared with the approaches mentioned above, the advantage of the mode shape-based method can be enumerated as follows. First, mode shape is more sensitive to local damages and can be used directly for multiple damage detection. Second, the mode shapes are less vulnerable to environmental effects, such as noise and temperature.[9] Due to these advantages, the mode shape surfaces as the aptest choice among various dynamic structural properties when it comes to health monitoring.

Conventional modal analysis methods can be broadly categorized into three major classes—the theoretical modal analysis (TMA), the experimental modal analysis (EMA), and operational modal analysis (OMA). TMA, also known as direct methods, investigate dynamic structural properties based on the mass and the stiffness matrix. On the other hand, EMA records the output response of vibrating structures subject to input excitation. The output response functions like impulse response function (IRF) and the frequency response function (FRF), obtained from field experiments, manifest the dynamic behavior of the structure. These functions help to calculate modal attributes such as mode shapes, modal frequencies, and other modal parameters. However, the measurement of theses two functions tends difficult for large structures. Besides, at times, the EMA is unable to accurately simulate both the real-world applications and the involved boundary conditions.

To address the problems accompanying the TMA and EMA, the operational modal analysis, also known as output-only analysis or ambient excitation, is established and widely used in applications involving towers, buildings, bridges, and off-shore platforms.[10–13] OMA utilizes the response, from structures operating in their ambient natural environment, to estimate the modal parameters. OMA is both fast to conduct and cost-effective. It encompasses the dynamic characteristics of the entire system rather than focusing on a few of the individual components, thus generating a more comprehensive representation of the working points. In OMA, the real-time loading features get linearized due to the involved random broad-band excitations. OMA efficiently handles the repeated modes or even the closed-spaced modes, rendering itself suitable for complex structures operating in their natural environment. Tremor-based health monitoring and structural control also use OMA. OMA requires dynamic measurements from physically-attached wired or wireless sensors (such as accelerometers).[14–17] Andreas et al[18] reported a multiplexed sensor array of fiber Bragg gratings (FBGs) as a quasi-distributed sensor to capture the mode shapes for beams. Based on a high-speed demodulator and a fast computation algorithm, the proposed method can determine the mode shapes. Sensor displacement techniques had always played a vital role in enhancing the quality of the captured modal analysis data, especially the mode shape. In previous studies,[19–22] the authors outlined different algorithms such as distributed wolf algorithm, improved artificial bee colony (IABC) algorithm, genetic algorithm, and multiobjective genetic algorithm to optimize the location and number of sensors on different structures like bridge and beam.

However, OMA with physically attached sensors has its limitations. First, the sensors' weight could result in mass-loading on lightweight structures that ultimately alter the structure's dynamics. Second, the spatial resolution of sensor-metric data collection is low as the sensors themselves are placed in a sparse grid over the whole structure under monitoring. The low spatial resolution of the sensor critically limits the accuracy of mode shape measurements and hinders the precise gauging of the associated dynamic properties. Finally, sensors' installation is a time consuming and labor-intensive process. Installation errors also influence the data sampling and prediction accuracy that requires specialized pre-processing of the acquired data.

Non-contact vibration measurement techniques attempt to address issues associated with contact-type measurement techniques. Huang et al[23] put forward the optical system called the AF-ESPI method, where the out-of-plane displacement estimation technique is employed to investigate the vibrational behavior of square-shaped isotropic plates. Ruan et al[24] introduced a vibration displacement measurement system based on the photoelectric method, which consists of a laser source, a linear charge-coupled device, and a corresponding software platform. The proposed method has been successfully applied to the multi-location displacement measurement. Even though these methods are non-contact and have relatively high accuracy, the overall equipment setup requires the deployment of lasers and is quite complicated. The cost of the complete equipment setup is pretty high due to the precision parts like the mirrors and filters, which makes it difficult for the large-scale application. These systems also require seismic isolation and a relatively quiet ambiance that makes them less robust to the noisy data. Schajer et al[25] used a similar technique called electronic speckle pattern interferometry (ESPI) to measure the vibration mode shapes of circular and band saws. It avoids the need to spread the powder over the saw blade surface and can identify low-frequency vibrations. However, the main disadvantages of ESPI are the cost and complexity of the equipment required and the need to color the target surface with reflective paint. Besides, the entire process is time-consuming. Improving upon ESPI, continuous scan laser Doppler vibrometry (CSLDV) came as a solution for mode shape measurement of the beams and the wind turbines.[26–28] Nonetheless, the primary shortcoming of CSLDV is the occurrence of speckle-noise during the high-speed laser scanning of the target surface. Speckle noise changes the intensity pattern of the laser light and adversely affects the measurement accuracy.

Recently, vision-based methods are gaining popularity. These methods make use of techniques like digital image correlation (DIC), pattern matching, and optical flow for the vibration displacement or the mode shape estimation task.[29–33] Feng et al[34] developed a novel non-contact vision-based method, utilizing a camera, for measuring simultaneous multipoint displacements. The method comprises of two different subpixel template matching techniques named upsampled cross-correlation (UCC) and orientation code matching (OCM). This method demonstrated high robustness, while extracting the local substructural displacements, even in a hostile environment. Nevertheless, it still requires significant pre-processing, such as cross-correlation calculations and peaks search. Significant pre-processing aggravates the involved complexity in generating useful results.

Compared with the conventional vision-based SHM method, the computer vision-based deep learning processing approach offers a new channel for excavating the massive data from an SHM system towards autonomous, accurate and robust processing of the monitoring data.[35] Kohiyama et al[36] utilize support vector machine (SVM) and deep neural network (DNN) models to classify the structural damage patterns. The DNN is first used for the data feature extraction from the input data, and then SVM can realize different unlearned damages patterns based on the features from DNN. Besides the DNN, CNN has demonstrated superb data abstraction capabilities in the SHM domain. Tang and his colleague[37] build up a CNN-based structural anomaly detection, which can learn the time and frequency domain features of the raw SHM data. The original data are transformed into the image format and then the CNN model is applied for the visual feature extraction. Based on these features, the CNN model can determine the class of the defect. Similarly, Khodabandehlou et al.[38] apply the 2D CNN model to predict the predefined damage states with recorded (acceleration) vibration response data from the actual highway bridge. The proposed CNN model can achieve a 100% classification accuracy for four different kinds of damage, which verifies the efficiency of the CNN model in the SHM domain. Besides the image level defect detection problem (classification), the CNN can also be utilized to solve the crack detection[39] and semantic segmentation problem. [40]

Besides the single deep learning model, recently, the hybrid deep learning model involving two architectures such as the CNN-LSTM is already in use in the domains of action recognition, text generation,[41,42] speech recognition, and sentiment analysis.[43–45] The reason for choosing the CNNs is its ability to automatically select useful features, whereas

LSTMs demonstrate superior learning ability from the sequential data. Xu et al[46] use CNN-LSTM architecture for face anti-spoofing by learning temporal features from the different videos dataset, which can extract features locally and densely as well as exploring the temporal structure from a continuous video stream. CNN-LSTM model has also been used for modal frequency identification problems.[47] The qualities mentioned above of both the CNNs and LSTMs make them the ideal candidates for modeling the spatio-temporal dependencies pervasive in the vibrational analysis of a structure. To train the deep learning architecture, we deemed mode-shapes as the video stream features that the CNN-LSTM model could recognize. Compared with other modalities like sensor data, the equipment and process for video stream acquirement are cheaper and more accessible. The video stream captured by a regular camera or high-speed camera is enough for further modal analysis. To accurately estimate the structural mode shape variations for SHM, we propose and outline, in this paper, a novel full-field computer vision-based modal shape analysis method. This method does not need structural surface preparation or image pre-processing and can be implemented relatively efficiently and autonomously. Also, each pixel of frames from the vibration video stream acts as a sensor capturing useful information for mode shape prediction. Thus, hundreds of thousands of pixel-sensors ensure the excellent performance of the proposed computer vision-based method, which is a considerable improvement compared with discrete physically attached sensors. The proposed computer vision-based mode shape prediction method requires only a camera, which further simplifies the inspection process, reduces the inspection cost, and thus ensures high precision at the same time. Specifically, we introduce a CNN-LSTM (convolutional neural network, long short-term memory) computer vision-based non-contact vibration measurement technique for automated mode shape prediction. In addition, two comparison methods called CNN- recurrent neural network (RNN),[48] and CNN-gated recurrent unit (GRU),[49] which have been used for the time-sequential data analysis, are selected to verify the superiority of the CNN-LSTM model for mode shape determination task.

This exploration endeavors to improve the intricate and lumbering vision-based techniques through a CNN-LSTM based AI approach. Primarily, we present an all-encompassing computer vision-based deep learning technique that discards any image pre-processing necessities and straightforwardly crumbles the vibration video outlines into fundamental mode shapes. Our deep learning computational pipeline is increasingly self-governing and displayed better execution when contrasted with recently referenced vision-based strategies.

The main contributions of this paper are as follows:

1. A state-of-the-art computer vision-based deep learning architecture that enables the fundamental mode shapes the perception of a vibrating structure.
2. A unique and non-existing dataset comprising auxiliary vibration recordings and their corresponding FEA-based modal displacements has been created.
3. The devised and outlined model performs proficiently, and the empirical outcomes are witness to its significant levels of extrapolation precision on an unseen dataset. Compared with two comparison methods CNN-RNN and CNN-GRU, the proposed CNN-LSTM performs better on mode shape prediction.

This paper is structured as follows: In Section 2, we discuss the process of visual vibration information collection that forms the basis for mode shape dataset generation. Section 3 outlines the main components of CNN-LSTM architecture that forms the backbone of our computational pipeline. Section 4 put forward the metrics for evaluating the architecture performance. Section 5 presents a formal and critical investigation of the obtained results. Finally, the main conclusions of our work are outlined, along with future avenues of related research.

2 | DATA COLLECTION

2.1 | Beam vibration frame generation

A shaker-based sweep test was performed to experimentally determine the mode shape of the test specimen in a more controlled setting to avoid noise in the data and incorrect data acquisition. The experiment was conducted with the help of a controller, shaker, and laser vibrometer (Figure 1).

Six different cantilever beam (continuous-type) structures with different materials and dimensions were used to collect data. The experimental specimen set consists of two aluminum beams, two copper beams, one brass beam, and one steel beam (Figure 2). Dimension, density, and Young's modulus of the six specimens are tabulated in Table 1. Basler's

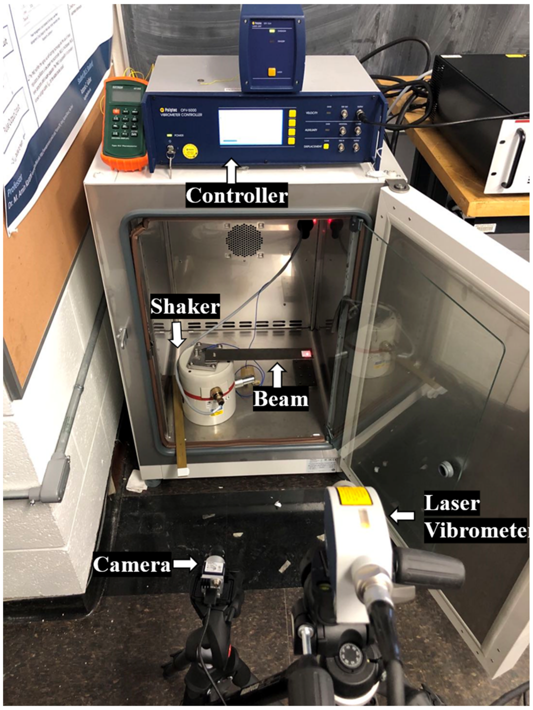

FIGURE 1 Vibration data collection setup

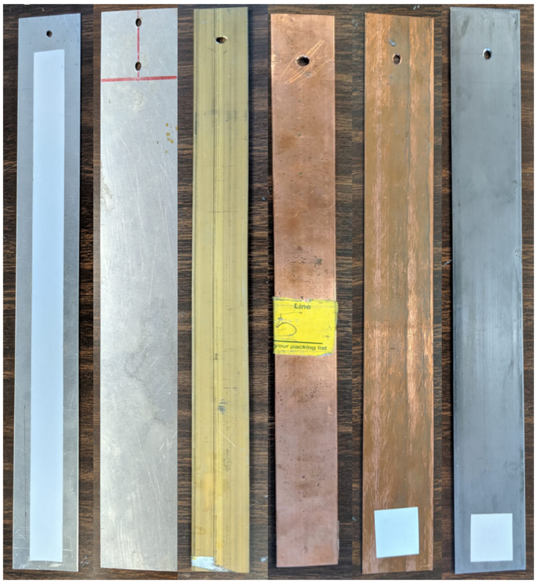

FIGURE 2 Six beam samples for data collection. From left to right: aluminum-long, aluminum-standard, brass, copper-narrow, copper-wide, and steel

high-speed camera was used to perform video measurements of the structure at a frame rate of 200 frames per second. For each specimen, we recorded five different vibration videos. The length of each video is around 3 min.

To additionally test the robust fitness of the CNN-LSTM model, another set of vibration videos, from two different viewpoints, was recorded for the six beam specimens. We trained our deep learning model on the video frames for the front viewpoint and then tested the same on top viewpoint data. The experimental setup for the top viewpoint is shown in Figure 3.

TABLE 1 The parameter of six cantilever beams

Material	Young's modulus (GPa)	Density (kg/m^3)	Dimension (mm) (L; W; T)
Aluminum-long	2700	69	354; 50.2; 2.57
Aluminum-standard	2700	69	273; 63.5; 1.27
Brass	8730	97	406; 25; 3.46
Copper-narrow	8960	128	305; 25.5; 3.31
Copper-wide	8960	128	305; 50.8; 1.8
Steel	7850	200	305; 50.75; 3.46

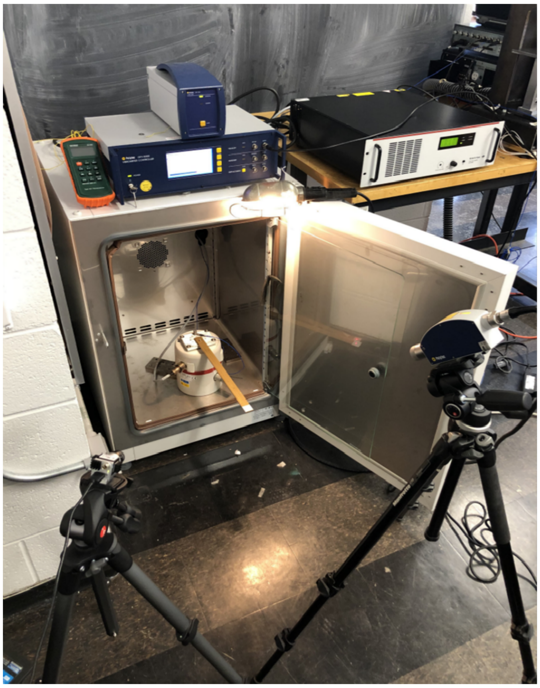

FIGURE 3 The new top view vibration experiment setup

The length of the video segment that contains perceptible vibration in each raw video is about 90 seconds. We utilize the OpenCV library[50] to extract frames of beam vibration from each video. The size of each raw frame is 728*544*3. Before feeding these raw frames into our proposed CNN-LSTM model, all raw frames were re-sized into 64*64*3 as the input data. In total, we select 88,800 frames from corrected dataset. Each consecutive set of 200 frames forms one sequential input data for the fundamental mode shape regression task. On the whole, there are 444 data samples in the data set. We split the entire dataset into 306 training samples, 48 validation samples, and 90 test samples. The vibration frames of each beam come from the five different videos of the same beam, prone to various combinations of the vacillating lighting conditions, vibration amplitudes, and other attributes such as reflected light streams, heat and sensor illumination etc. The distance between the camera and the beam and illumination conditions in each of the videos are slightly different. Varying recording conditions amount to what we can refer to as the additional noise in the training dataset. Figure 4 displays the image frames from two different videos of the same steel beam.

For the top viewpoint vibration dataset, we collect 3000 frames for each beam resulting in a total of 18000 frames (for six beams) from the top view experiment. These frames were used to test the trained model for its robustness on viewpoint change. The vantage point for the top viewpoint is depicted in Figure 5. The ModeShape database is located at GitHub.[51]

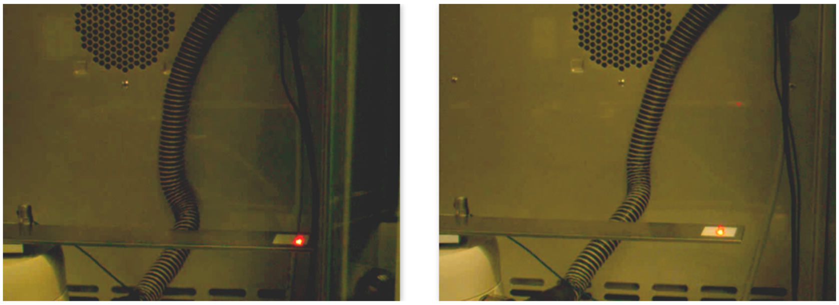

FIGURE 4 The two frames of steel vibration from two different front view videos

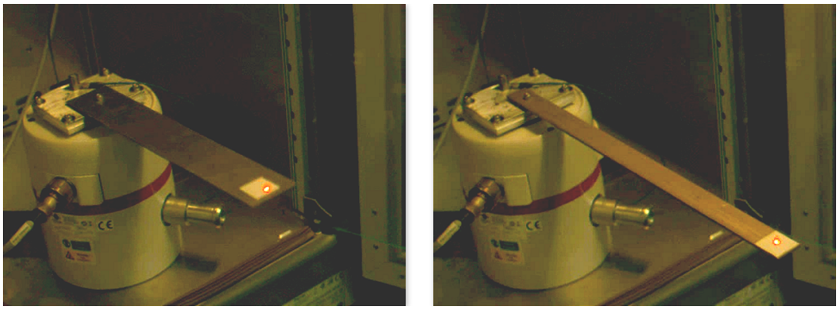

FIGURE 5 The top view of specimen beam vibration frames

2.2 | Mode shape label generation

As a supervised learning architecture, our model *learns* a mapping between the input video-frames and the corresponding output labels, that is, f_{Model}: $[X_{Video - frames}] \rightarrow Y_{ModeShape}$. Herein, we generate the output labels, the point-wise displacements for the respective mode shapes, using FEA-based methodology. Although the physical experiments help to characterize the dynamic behavior of a structure in terms of its modes of vibration, they have certain disadvantages associated with them. First of all, it is difficult to accurately measure the particle displacement using a shaker based experimental setup. We can measure specific indirect attributes, such as frequency response functions (FRFs), related to mode shapes. However, converting these parameters to actual mode shapes is in itself a long haul. Second, the deployment of a vibration shaker for gigantic outdoor specimens is not practically feasible. Finite element analysis (FEA) provides a solution to the issues, as mentioned above, related to the experimental vibration setup.

Although FEA is a numerical modeling approach, contrary to the sensor-based approach of the vibration shaker experiment, the FEA comes with a set of specific advantages which makes up for the involved assumptions. First of all, it outputs an overall results set demonstrating the physical response of the system, to the input vibration stimulus, at any location. Many of these physical responses get ignored in the actual physical or analytical approach owing to the system complexities. Second, FEA provides an option to execute safe simulation for potentially destructive or impractical vibrating conditions and failure modes. Third, the FEA model can be used to extrapolate the actual experimental results. Additionally, the FEA model ensures efficiency on the economic front. Hence, in this paper, the FEA model serves as a base-line model for generating the training labels, that is, the point-wise modal displacements and also for validating the outputs of the proposed deep learning model.

The first three eigenfrequencies and mode shapes are derived by eigenfrequency study in COMSOL multiphysics based FEA. The physics used in the model is the solid mechanics module under the structural mechanics component that provides a range of equations for specifying subdomains, boundaries, edges, and points. The materials are linear elastic with the isotropic model that has mechanical properties, as specified in Table 1. The boundary condition defined for each beam is clamped-free. The clamping side has fixed constraints with zero displacements in all directions in the selected boundaries. The boundary condition for each beam is defined as a fixed constraint for the clamp side of the beam, which means the displacement is zero in all directions in the selected boundaries (Figure 6). Mode shapes

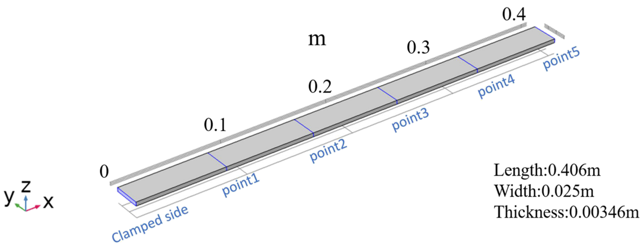

FIGURE 6 The boundary condition and five sample points on the brass beam

provide the relative position of the points of the structure concerning each other. Another frequency study, for absolute deflections, is performed by exciting the structure under each of natural frequencies. For analysis, we divided each beam into five equal proportions along the length of the beam, resulting in six sample points on the beam. Due to one fixed end on each beam, for each mode shape, we obtain displacement values of five discrete points (Figure 6).

In the frequency domain, acceleration and displacement are related to each other by following equation

$$a_{base} = (2 * \pi * f)^2 * u_{base} \tag{1}$$

where a_{base}, f, and u_{base} are base acceleration, frequency, and base deflection, respectively. From Equation (1), we get the base deflection and add that to relative deflection coming from the FEA model. The governing equations of the FEA model for linear elastic materials in the eigenfrequency study are defined as

$$-\rho\omega^2 u = \nabla.S \tag{2}$$

$$-i\omega = \lambda \tag{3}$$

$$S = C:\epsilon \tag{4}$$

$$C = C(E, \upsilon) \tag{5}$$

$$\epsilon = 1/2[\nabla u^T + \nabla u] \tag{6}$$

where ρ, u, ω, S, λ, E, υ, and ϵ are density, displacement field, angular frequency, stress, eigenvalue, Young's modulus, Poisson's ratio, and strain, respectively.

We performed the finite element analysis (FEA) to characterize the structural dynamics by constructing a numerical model. To compensate for the assumptions made during the numerical modeling, we compared the frequency response functions (FRFs) as obtained from both the FEA and the experimental shaker-based experiment. Closer proximity in results, from both the experiment and the FEA, substantiates the accuracy of the FEA in modeling the vibrational dynamics. Figure 7 illustrates one such comparative analysis between the frequency response functions for different beam specimen, as obtained from both the FEA and the experimental setup.

One possible source of minor deviations, as depicted in Figure 7, between the results of physical experiments and the FEA is that the boundary conditions might be slightly different. For example, the flexibility of the clamped side of the beam mounted on the shaker is assumed as rigid in FEA, which may not be completely rigid in the physical experiment. Additionally, possible imperfections present in the beam can also affect the results in the physical experiment.

FEA results from the COMSOL software determines the absolute displacement values of the sample points used for labeling and training purposes. Figure 8 shows an example of the FEA-based mode shape for the brass specimen beam. As mentioned earlier, five points were monitored continuously for different orders of mode shape. In total, for each specimen, the corresponding training labels contain 15 displacement values for 3 mode shapes (5 values per mode), as displayed in Table 2.

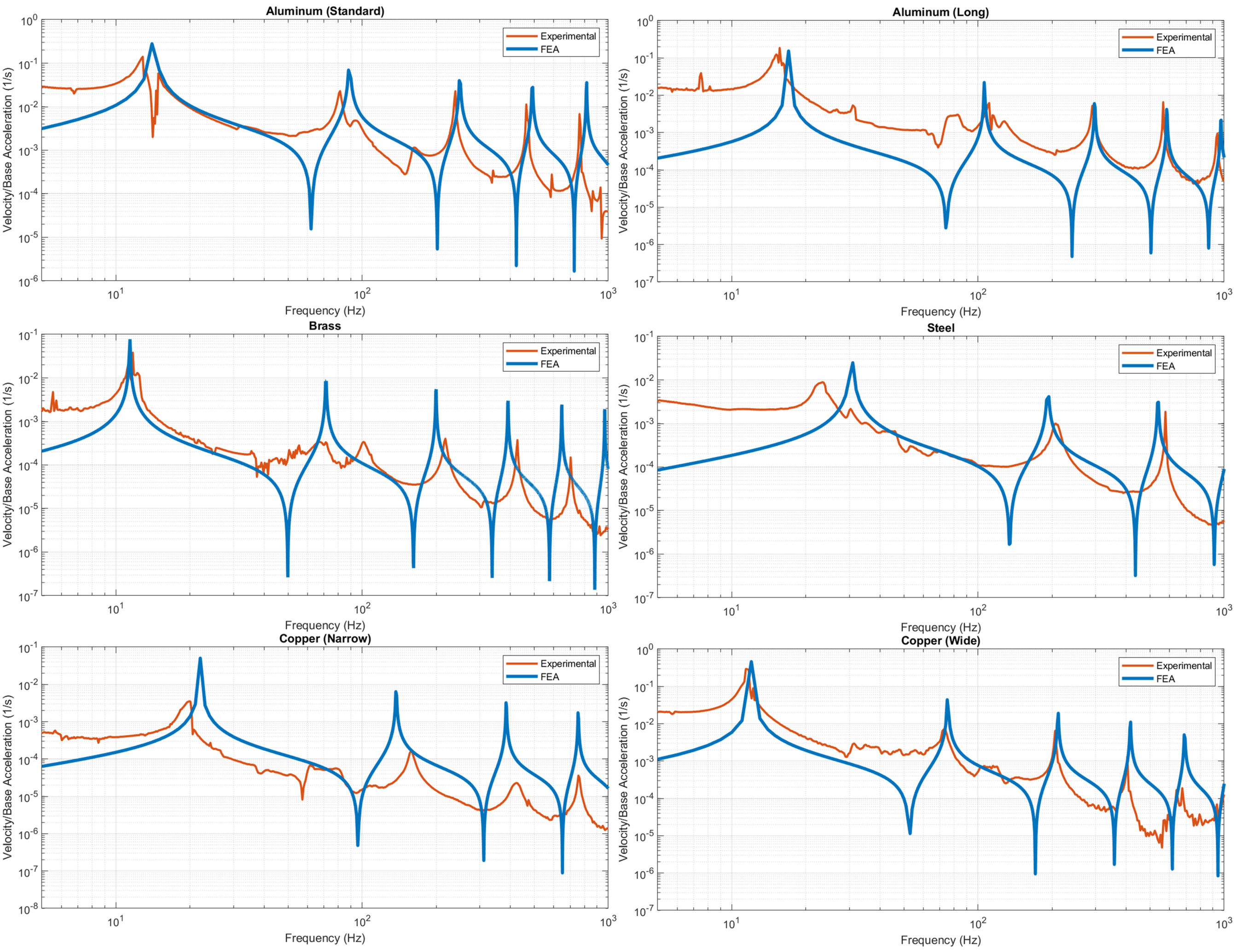

FIGURE 7 Comparison of velocity frequency response functions (FRF) between experiment and FEA, from top left: AL standard, AL long, brass, steel, copper narrow, and copper wide

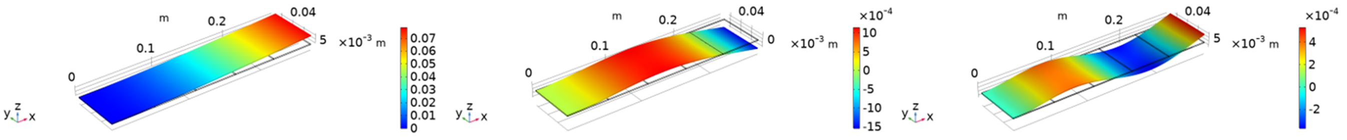

FIGURE 8 The displacement point sampling of aluminum-standard beam

TABLE 2 The absolute mode shape (mm) of all cantilever beams

Material/Specimen	First mode shape	Second mode shape	Third mode shape
Aluminum-long	1.25, 3.51, 6.69, 10.36, 14.03	3.96, 9.31, 8.09, −0.87, −13.41	8.15, 7.39, −6.42, −5.67, 13.40
Aluminum-standard	1.89, 3.75, 6.38, 9.39, 12.49	3.31, 7.89, 6.86, −0.62, −11.12	6.86, 6.11, −5.39, −4.98, 10.94
Brass	1.96, 4.27, 7.68, 11.61, 15.49	4.35, 9.92, 8.65, −0.97, −14.41	8.79, 7.78, −6.74, −5.78, 14.60
Copper-narrow	1.01, 2.82, 5.41, 8.45, 11.40	3.29, 7.67, 6.63, −0.74, −11.01	6.67, 5.95, −5.30, −4.48, 11.15
Copper-wide	0.89, 2.90, 5.62, 8.81, 12.14	3.50, 8.10, 7.25, −0.65, −11.76	7.26, 6.64, −5.75, −5.08, 11.85
Steel	0.99, 2.92, 5.75, 8.90, 12.25	3.60, 8.25, 7.22, −0.766, −11.88	7.19, 6.39, −5.77, −5.01, 11.87

3 | ARCHITECTURE DESCRIPTION

In general, image processing techniques are computationally expensive and require non-trivial image transformations. What is needed are efficient computational pipelines to extract the mode shapes of the vibrating structures. In this section, we introduce a CNN-LSTM based deep learning architecture that exploits the time series dependency among the frames of the video acquired through a monocular camera to automate the entire mode shape identification process. The outlined CNN-LSTM method alleviates the need for conventional contact sensors while achieving significant levels of accuracy. The overall computational pipeline is depicted in Figure 9.

3.1 | CNN

The convolutional neural network (CNN) has shown exemplary performance in numerous computer vision and pattern recognition problems. The role of CNN in our architecture is to identify the structural features in the regular lattice of pixels. Our proposed CNN model comprises alternate layers of convolution and max-pooling, followed by a fully connected layer. Attributes determining the performance of the CNN model are feature maps, kernel size, and spatial strides. The convolution layer consists of several neurons, and each of them acts as a convolution kernel. These kernels work by dividing the vibration frame image into smaller blocks to extract motif features. We deploy a rectified linear unit (ReLU) activation function to map the non-linearity between the inputs and outputs of the CNN model. The ReLU activation is applied just after the convolution layer of the CNN model and can be expressed as

$$ReLU(x) = max(x,0) \tag{7}$$

This activation function helps to process the information gathered by the convolution operator. After convolution comes to the max-pooling layer. Max-pooling layer helps to downsample the image and extricate relatively dominant features from the field neighborhood. Pooling operation helps to reduce the size of feature-map, thus reducing the network complexity and also improving the generalizability of the model. After the last pooling layer, we have a flatten layer that transforms the image tensor into a vector of size equal to that of the number of elements in the input tensor. As one training sample contains 200 frames, so each time, 200 images are fed into the CNN model with 3*3 kernel size and 64 feature maps. After convolution and max poling operations with 2*2 pooling size, each image gets converted to

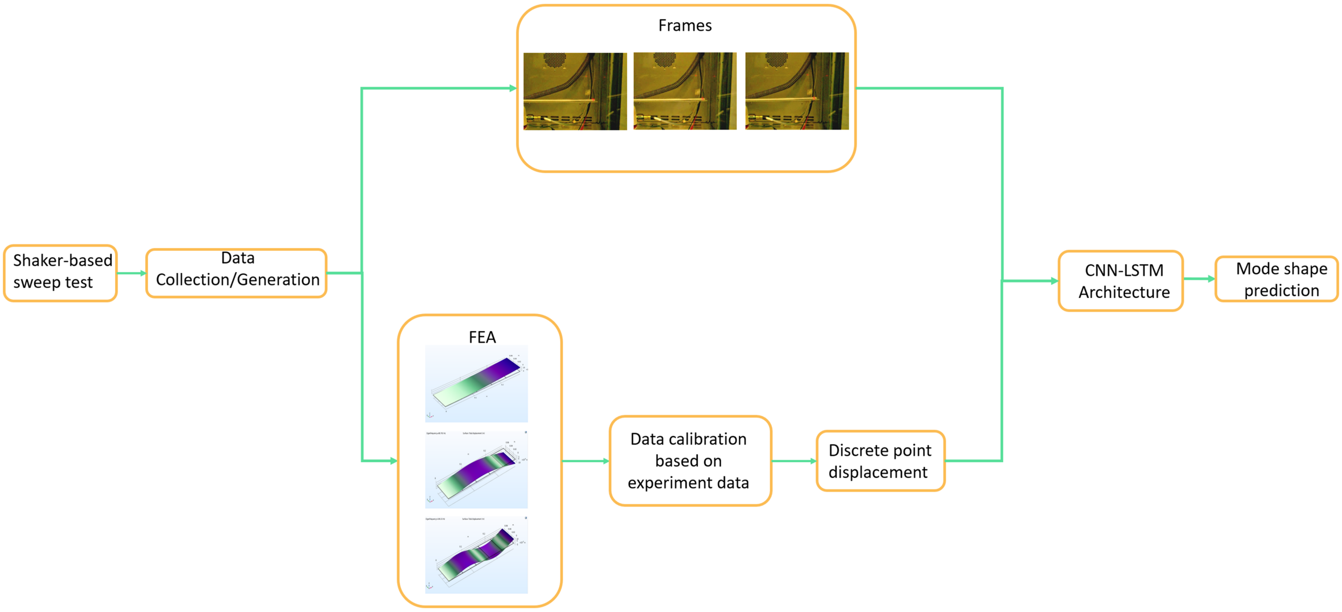
Frames
Shaker-based sweep test
Data Collection/Generation
FEA
Data calibration based on experiment data
Discrete point displacement
CNN-LSTM Architecture
Mode shape prediction

FIGURE 9 Computational pipeline

a feature vector of size 1*4096. Two hundred different feature vectors are sent to the LSTM layer with specific hidden units to predict the first three mode shapes. The used CNN architecture and associated hyperparameters are depicted in Figure 10.

3.2 | LSTM

LSTMs have emerged as one of the most effective tools for processing sequential datasets. Due to their recurrent processing capabilities, they have been used to solve the state-of-the-art problems in domains of acoustic speech modeling, video analysis, and audio synthesis. Another essential feature of the LSTM architecture block is its information retention ability. Each LSTM cell has a memory unit that assists in maintaining its state over time. Besides, few gating units regulate inflow and outflow of information within the cell.

The overall CNN-LSTM architecture has the CNN and the LSTM blocks interacting periodically. Herein we feed the output, tensor imbibing the pixel information, of the CNN model into the LSTM cells in sequential order. If x_t is one of the CNN model outputs at any time step t, the LSTM cell will take that as input for that particular time instant, process it, and generate the hidden activations a^t represented by Equation (10). This activation value goes through a series of recurrent processing steps to generate the final output and thus forming a memory flow over time that aids in modeling long term Spatio-temporal dependencies present in a sequence of video frames. Each LSTM cell consists of a latent cell state c^t, calculated using Equation (9), which serves as a memory and helps hidden units a^t in retaining information from the past. We use $c^{\sim t}$ as a placeholder, using Equation (8), to initially replace c^t. The memory state c^t is generated by combining c^{t-1}, a^{t-1}, and the input features at time step t. Figure 11 showcases the LSTM cell. LSTM cell takes the input x_t from

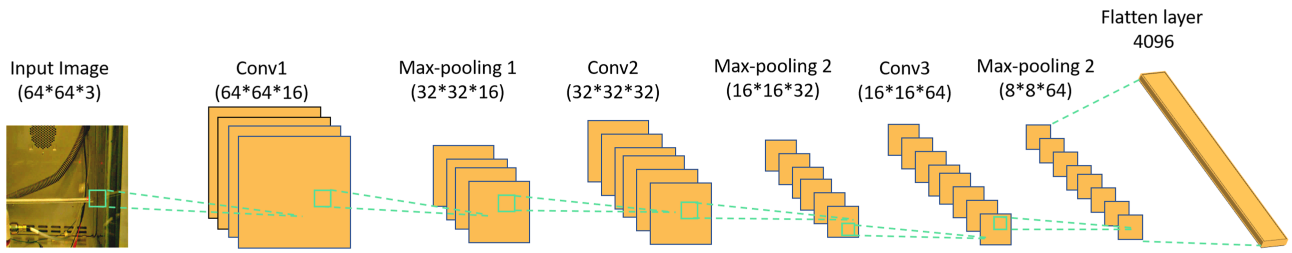

FIGURE 10 The architecture of the CNN part for the proposed architecture

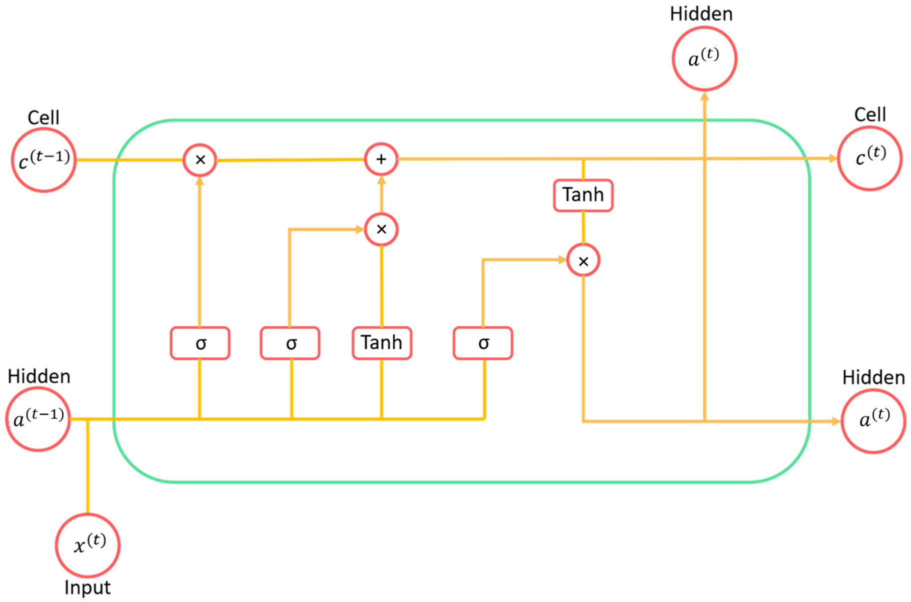

FIGURE 11 The architecture of the LSTM cell

the current step and the inherited information a^{t-1} from the previous time step to generate the cell activation a^{t}. Since a^{t} is the total of information from all the previous time steps, it is pivotal in determining the final output.

$$c^{\sim t} = tanh(W_a^c a^{t-1} + W_x^c x_t) \tag{8}$$

$$c^t = f^t \otimes c^{t-1} + u^t \otimes c^{\sim t} \tag{9}$$

$$a^t = o^t \otimes tanh(c^t). \tag{10}$$

Here, we use the weight parameters, W_a^c and W_x^c, to generate candidate cell state and pretermit the bias terms as they get absorbed into weight matrices. Thereafter, we introduce a forget gate layer f^t, an update gate layer u^t, and an output gate layer o^t, as

$$f^t = \sigma(W_a^f a^{t-1} + W_x^f x_t) \tag{11}$$

$$u^t = \sigma(W_a^u a^{t-1} + W_x^u x_t) \tag{12}$$

$$o^t = \sigma(W_a^o a^{t-1} + W_x^o x_t) \tag{13}$$

All the input video frames from sample data go through the abovementioned processing order. The final output is the 15 point displacements for three different mode shapes. Figure 12 showcases the architecture of the proposed CNN-LSTM model.

4 | EVALUATION METRICS

4.1 | Generalizability

The generalizability refers to the ability of a learned model to fit an unseen instance within its training input domain range. In this paper, we use MSE (mean squared error) as the evaluation criteria for generalizability. MSE is the quantification of the average squared errors between the prediction and ground truth values.

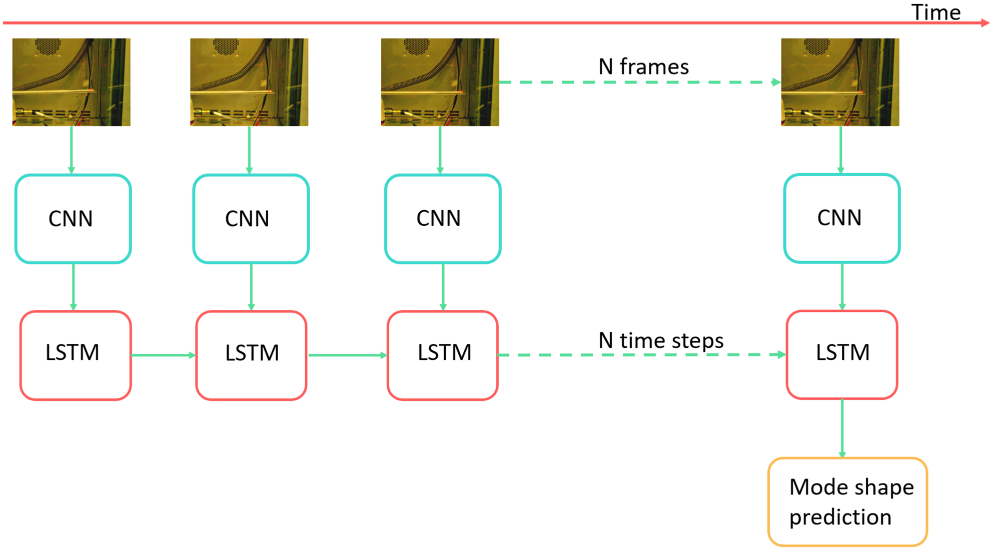

FIGURE 12 The architecture of the CNN-LSTM model

The MSE can be expressed by Equation (14):

$$MSE = \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n} \tag{14}$$

where $\hat{y}$ is the predicted value, y is the ground truth value, and n is the sample size.

Here, the MSE value between predicted mode shape and real mode shape in three different natural frequencies for six different beams is chosen to evaluate the model performance.

4.2 | Robustness on noisy data

Robustness appraises a model's performance in the presence of noisy data and unstructured state space. Higher the robustness, the greater is the model's capability to accomplish even in the presence of imprecise measurements. A robust model proves to be efficacious in real-life situations where, most of the time, we have a corrupt dataset at hand.

Vacillating lighting conditions, vibration amplitudes, and other attributes such as reflected light streams, heat, and sensor illumination constitute what is known as noisy data. Vibration videos of the same beam are prone to various combinations of the above mentioned noisy experimental conditions that often tends to taint the training data with noise. We create a comprehensive training dataset, comprising vibration videos of six different sample beams. The distance between camera and beam and illumination conditions in each of the videos are slightly different. Figure 4 displays the images from two different videos of the steel beam.

Table 1 lists the six specimen beams used to train the model for testing its robustness on the noisy data. We train the model with three recordings comparing to every one of the six beams and test them, for the mode shape expectation task, on the remaining two recordings of the individual beams.

4.3 | Extrapolability

Extrapolability is the measure of performance of a model beyond the range of initial training data. In other words, extrapolability is the model's prescient capability to predict accurately on data not seen by the model during the training stage. Unlike the robustness to noise metric, the extrapolability metric measures the ability of the outlined CNN-LSTM model to predict the mode shapes of a test specimen that is not part of initial training data.

To compute the extrapolability metric, we train the CNN-LSTM model on data samples obtained from five out of six beams and use the trained model for predicting the mode shape of the sixth beam (excluded information that was not utilized during the training of the model). In total, we performed six different examinations—each trial comprise training on data samples acquired from five beams, trailed by testing on the remaining one.

4.4 | Robustness on viewpoint change of video

Changing the viewpoint of the camera can change the background, foreground, and the size and appearance of the test specimen observed through the camera. It is therefore essential to measure the robustness of the CNN-LSTM model in the presence of camera viewpoint changes. The two different viewpoints of six different beams vibration frames have been discussed in Section 2. The model is trained on front view video frames and tested on the top view frames.

4.5 | Computational resources

The network is trained using Keras (Tensorflow GPU 1.14.0 backend), CUDA 10.0 toolkit, and cuDNN 7.0 support in a machine of Alienware R8 desktop, which has 16GB RAM and a 8 GB video RAM RTX 2080 super GPU.

4.6 | Hyperparameter tuning

Hyperparameter tuning refers to the process of determining ideal or optimal values of hyperparameters (knobs for tuning and enhancing the performance) for a learning algorithm.[52] LSTM layer nodes, learning rate, number of CNN layers/nodes, and batch size as principle hyperparameters in our CNN-LSTM model. Moreover, early stopping calculations are applied during the preparation process to avoid the over-fitting issue (when the trained model tries to predict a trend in data that is too noisy). We utilize the L2 regularization and early-stopping calculations to avoid any overfitting. L2 regularization calculation works by applying penalties on layer parameters. L2 norm characterizes the regularization term as the aggregate of the squares of all the component loads.

4.6.1 | Tuning for robustness on noisy data

Various combinations of hyperparameter values were utilized to prepare our CNN-LSTM model for the mode shape prediction in three different regular frequencies. The box plot of MSE values, over the test set, for five of the chosen models are shown in Figure 13. Table 3 lists the details of hyper-parameter sensitivity analysis for five different models. Among the five models, the fifth model obtains the best performance with an MSE value of 0.0049 on the test dataset.

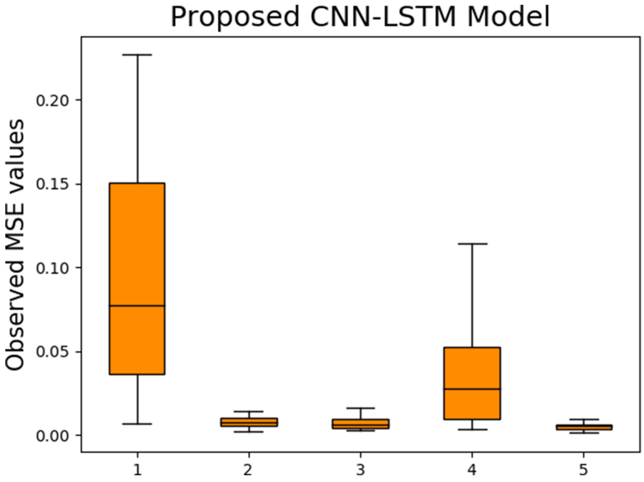

FIGURE 13 The box plot of MSE value for five separate CNN-LSTM models with different hyperparameters

TABLE 3 Hyper-parameters of CNN-LSTM architecture for robustness on noisy data (k: kernel size, c: channel number, n: node number)

Configuration	1	2	3	4	5
Convolutional layer	k(3 × 3)/c(16)	k(3 × 3)/c(16)	k(3 × 3)/c(16)	k(3 × 3)/c(16)	k(5 × 5)/c(16)
Max pooling layer	k(2 × 2)	k(2 × 2)	k(2 × 2)	k(2 × 2)	k(2 × 2)
Convolutional layer	k(3 × 3)/c(32)	k(3 × 3)/c(32)	k(3 × 3)/c(32)	k(3 × 3)/c(32)	k(5 × 5)/c(32)
Max pooling layer	k(2 × 2)	k(2 × 2)	k(2 × 2)	k(2 × 2)	k(2 × 2)
Convolutional layer	k(3 × 3)/c(64)	k(3 × 3)/c(64)	k(3 × 3)/c(64)	k(3 × 3)/c(64)	k(5 × 5)/c(64)
Max pooling layer	k(2 × 2)	k(2 × 2)	k(2 × 2)	k(2 × 2)	k(2 × 2)
LSTM layer	n(60)	n(50)	n(30)	n(50)	n(30)
Dense layer	c(15)	c(15)	c(15)	c(15)	c(15)
Learning rate	0.001	0.001	0.001	0.01	0.001
Batch size	10	5	10	5	10

The batch size and learning rate are set to 10 and 0.001, respectively. The whole training process involving 100 epochs takes about 1.22 h, and the training and validation loss is 3.6×10^{-6} and 5.12×10^{-5}.

Besides the hyper-parameters mentioned above, the image size is another significant parameter that affects the model's prediction performance over the test dataset. The smaller image can realize the time and memory-efficient training process, which may affect the prediction accuracy due to the feature limitation problem. On the contrary, the bigger image obtaining more valuable features will utilize higher computational resources and longer training time. There is a trade-off between prediction accuracy and computational cost. In this paper, the three different sizes of the images, including $40 \times 40 \times 3$, $64 \times 64 \times 3$, and $128 \times 128 \times 3$, are applied to select the best image size for the mode shape prediction problem. Due to the hardware limitation, the image size bigger than $128 \times 128 \times 3$ will cause the memory exhaust problem, so no image bigger than $128 \times 128 \times 3$ is considered in this test. The training time for each image size involves 100 epochs, and the bigger image size requires a longer training time. The trained model's performance for $64 \times 64 \times 3$ and $128 \times 128 \times 3$ is the same that the MSE value over the validation dataset is both 0.0013. However, the computational time for $128 \times 128 \times 3$ is around 5 times of $64 \times 64 \times 3$ image. Even though the smallest image size only needs 1 h for 100 epochs training, the performance is much worse than the other two images. Based on these comparisons, the image size for all training in this paper is selected as $64 \times 64 \times 3$ (Table 4).

4.6.2 | Tuning for extrapolability

The hyperparameter tuning process for computing extrapolation metrics, for the most part, centers around the learning rate, batch size, early stop calculation, and L2 regularization. The early-halting calculation utilized the "patience" parameter with a magnitude of 5. The incentive for kernel regularization and bias regularization are both set as 0.015 for L2 regularization. The batch size is set as 5, and the training epoch is set as 100.

4.6.3 | Tuning for robustness on viewpoint change

A distinct mix of the learning rate, L2 regularization values were tried on robustness to camera viewpoint change tasks. The value for kernel regularization and bias regularization are both set as 0.01 for L2 regularization. The batch size is set as ten, and the training epoch is set as 100. The "patience" parameter of the early-stop algorithm is set as 10 to solve the over-fitting problems, affecting the number of epochs and the whole training time.

5 | RESULTS

5.1 | Robustness on noisy data

Figure 14 shows the box plots of the CNN-LSTM model's MSE value for the robustness of noisy data tests. The mean value for MSE is around 0.008. A low value of MSE showcases the superior performance of our CNN-LSTM model for the mode shape determination task. Table 5 is the compilation of "robustness on noisy data" metric values for the three mode shapes. Table 6 shows the MSE values for each beam sample obtained by the CNN-LSTM model and two comparison methods. The CNN-RNN model gets the worst performance among the three models, especially for the copper-wide and steel beam. The average MSE value for all six beams of CNN-RNN is 1.028. The CNN-GRU model realizes better prediction accuracy than the CNN-RNN model for all six samples with an average MSE value of 0.097. Since

TABLE 4 The model performance for different input size

Input image size	Training time (h)	MSE value over validation dataset
$40 \times 40 \times 3$	1	0.004
$64 \times 64 \times 3$	1.22	0.0013
$128 \times 128 \times 3$	6.25	0.0013

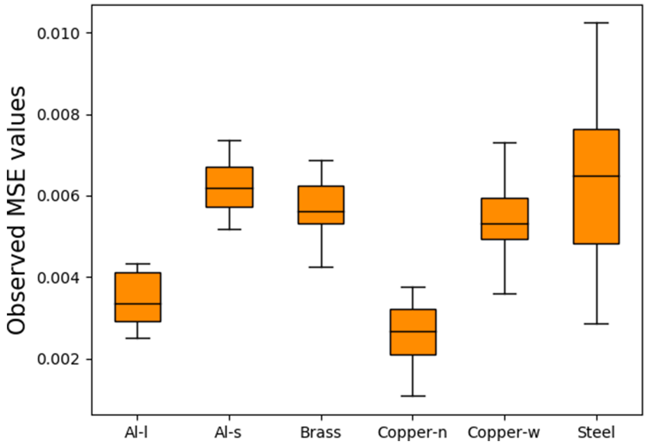

FIGURE 14 The box plot for the proposed CNN-LSTM model's ability of robustness for noisy dataset

TABLE 5 The mode shape results (mm) of robustness on noisy data

Material/specimen	First mode shape	Second mode shape	Third mode shape
Aluminum-long	1.25, 3.59, 6.78, 10.32, 13.95	4.04, 9.3, 8.06, −0.82, −13.45	8.16, 7.45, −6.37, −5.71, 13.42
Aluminum-standard	1.85, 3.75, 6.39, 9.28, 12.43	3.32, 7.93, 6.92, −0.61, −11.24	6.91, 6.28, −5.31, −4.99, 10.98
Brass	2, 4.39, 7.76, 11.57, 15.39	4.44, 9.96, 8.62, −0.89, −14.46	8.82, 7.82, −6.67, −5.89, 14.64
Copper-narrow	0.99, 2.87, 5.47, 8.43, 11.36	3.34, 7.7, 6.64, −0.75, −11.93	6.69, 6.03, −5.25, −4.53, 11.19
Copper-wide	0.93, 2.99, 5.57, 8.83, 12.08	3.58, 8.14, 7.29, −0.61, −11.84	7.42, 6.58, −5.77, −5.08, 11.93
Steel	0.98, 2.99, 5.86, 8.94, 12.23	3.68, 8.34, 7.22, −0.72, −12.01	7.18, 6.44, −5.73, −5.08, 11.97

TABLE 6 The MSE value for mode shape prediction of three models on noisy dataset

	Beam					
Model	Aluminum-long	Aluminum-standard	Brass	Copper-narrow	Copper-wide	Steel
CNN-RNN	0.057	0.046	0.028	0.037	0.37	0.49
CNN-GRU	0.012	0.017	0.0076	0.028	0.2	0.32
CNN-LSTM	0.004	0.006	0.006	0.0025	0.0054	0.0063

the LSTM and GRU can maintain information in the memory longer than the original RNN, the CNN-GRU, and CNN-LSTM can achieve better performance than the CNN-RNN model.

The first column in Figure 17 shows the mode state, represented by five distinct point displacements (absolute), of six beams in three distinctive regular frequencies generated by the outlined CNN-LSTM model. The depicted results (from top to bottom) are aluminum-long, aluminum-standard, brass, copper narrow, copper wide, and steel beam specimens. The red triangle line represents ground truth (FEA-based values for the three mode shapes) while the blue, cyan, and green dashed lines represent predicted values of the first, second, and third mode shapes, respectively. In every one of the six plots, on the left, these three prediction lines nearly trace the ground truth for five spatial points on the respective beams. The leftmost column of Figure 17 further demonstrate the efficacy of the CNN-LSTM model and also provides a possible reason behind a low MSE value of 0.004.

5.2 | Extrapolability

Table 7 tabulates the extrapolatability metric values of three basic mode shapes, outside the training domain range, for the six beams. Figure 15 depicts the MSE values of the best CNN-LSTM model as a box plot. Although the test dataset

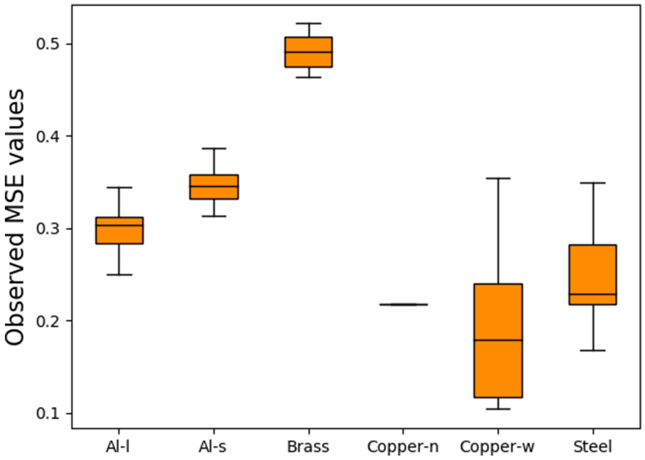

FIGURE 15 The box plot for the proposed CNN-LSTM model's extrapolability on six different beams

TABLE 7 The mode shape results (mm) of extrapolability

Material/specimen	First mode shape	Second mode shape	Third mode shape
Aluminum-long	1.57, 3.5, 6.47, 9.85, 13.36	3.82, 8.56, 7.57, −0.86, −12.91	7.25, 6.9, −6.02, −5.03, 12.44
Aluminum-standard	1.85, 4.15, 5.43, 8.55, 11.9	2.8, 7.47, 6.47, −0.61, −12.1	6.62, 5.78, −6.25, −5.59, 10.6
Brass	1.81, 3.72, 6.96, 10.94, 14.1	3.98, 9.69, 7.96, −0.9, −13.69	8.11, 7.1, −5.95, −5.21, 13.53
Copper-narrow	0.89, 2.94, 5.74, 9.09, 12.5	3.27, 8.03, 7.05, −1.19, −12.64	6.53, 5.85, −5.67, −4.91, 10.96
Copper-wide	1.36, 3.58, 5.99, 9.27, 12.79	3.69, 8.08, 7.08, −0.64, −12.67	7.37, 6.58, −5.67, −5.17, 12.11
Steel	0.63, 2.63, 5.5, 8.65, 12.1	2.89, 7.77, 6.51, −1.22, −12.67	6.89, 6.01, −6.34, −5.69, 11.46

TABLE 8 The MSE value for mode shape prediction of three models on extrapolability

	Beam					
Model	Aluminum-long	Aluminum-standard	Brass	Copper-narrow	Copper-wide	Steel
CNN-RNN	1.43	0.84	2.86	2.79	0.66	1.15
CNN-GRU	0.65	1.53	3.5	2.4	0.41	1.42
CNN-LSTM	0.31	0.35	0.51	0.21	0.19	0.24

is totally outside the training domain range, CNN-LSTM models perform well on the mode shape prediction task. Extrapolation results specified in Table 7 demonstrate the viability of the CNN-LSTM to model for mode shape extrapolation task. Table 8 illustrates the three model's performance over extrapolability dataset. The CNN-RNN and CNN-GRU model still get worse performance than the CNN-LSTM model due to the shorter period of information analysis ability that the average MSE value is about 1.63 and 1.65. The CNN-RNN and CNN-GRU model both obtain the best mode shape prediction value for a copper wide beam that the MSE value is 0.66 and 0.41 separately.

The middle column of Figure 17 diagrams the basic mode shapes as predicted, by the proposed CNN-LSTM model, on an extrapolated dataset. The predicted results (from top to bottom) are aluminum-long, aluminum-standard, brass, copper narrow, copper wide, and steel beam specimens. The red triangle line represents the ground truth (FEA-based values for the three mode shapes) while the blue, cyan, and green dashed lines represent predicted values of the first, second, and third mode shapes, respectively.

The MSE values for extrapolation tests are greater when contrasted to that of the robustness to noise tests. However, it is to be noted that in the robustness to noise tests, training models have access to data samples from the identical material beams. On the other hand, in extrapolation tests, the CNN-LSTM model, despite having no access to the same material data samples during training, has a very similar predictive ability to FEA. The best predictive values (an MSE value of 0.19) of the CNN-LSTM model were witnessed in the case of “copper wide” beam specimen, which is quite a decent performance for a data-driven model on an utterly unseen dataset. However, the performance for the “brass” beam specimen had room for improvement, where the MSE value surpasses the 0.51 mark. One experimental barrier that thwarted the model performance is the poor lighting conditions that existed in data collection experiments due to the cramped space requirements of the experimental setup. Poor lighting prevents the beam sample from being sufficiently illuminated and puts a greater onus on the CNN-LSTM model, fed with low-quality picture frames, to assimilate actionable information from the pixels.

Additionally, regardless of the camera speed being just 200 frames per second, which is less than the third-order natural frequency of the beams, our CNN-LSTM model shows quite a decent performance even on an unseen/novel specimen test dataset. This good prescient ability of the CNN-LSTM model can be attributed to the judicious design amalgamation of CNN and LSTM layers. The spatio-temporal nature of the CNN-LSTM architecture encourages it to extract useful information from the nearby sub-structural pixel arrangement and across different frames (time-steps), compensating for the inadequacies of the camera.

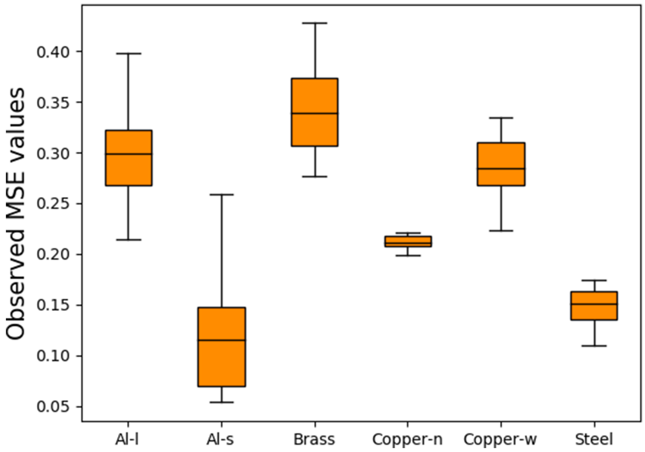

FIGURE 16 The box plot for the proposed CNN-LSTM model's robustness on viewpoint change

TABLE 9 The mode shape results (mm) on viewpoint change

Material/specimen	First mode shape	Second mode shape	Third mode shape
Aluminum-long	1.28, 3.47, 6.39, 10.03, 13.32	3.8, 8.5, 7.56, −0.85, −12.16	7.65, 6.9, −5.95, −5.36, 12.89
Aluminum-standard	1.63, 3.59, 6.13, 9.12, 12.33	3.27, 7.78, 6.81, −0.53, −11.25	6.9, 6.22, −5.29, −4.96, 11
Brass	2, 3.41, 6.92, 11.11, 14.85	4.11, 9.01, 7.83, −0.81, −13.63	8.51, 7.45, −6.91, −5.23, 13.91
Copper-narrow	0.98, 2.75, 5.36, 8.52, 11.29	2.86, 7.41, 6.62, −0.73, −12.68	5.95, 5.35, −5.77, −5.08, 10.26
Copper-wide	0.87, 3.35, 5.87, 9.1, 13.11	3.47, 8.7, 7.13, −0.88, −12.27	7.55, 6.13, −6.39, −5.47, 12.53
Steel	1.24, 3.35, 5.96, 9.88, 12.64	3.08, 8.3, 7.22, −0.72, −12.46	7.09, 6.37, −5.55, −4.99, 11.73

TABLE 10 The MSE value for mode shape prediction of three models on viewpoint change

	Beam					
Model	Aluminum-long	Aluminum-standard	Brass	Copper-narrow	Copper-wide	Steel
CNN-RNN	0.84	1.33	1.79	1.61	0.73	0.53
CNN-GRU	1.39	0.61	1.37	2.63	0.61	0.47
CNN-LSTM	0.29	0.11	0.34	0.22	0.28	0.15

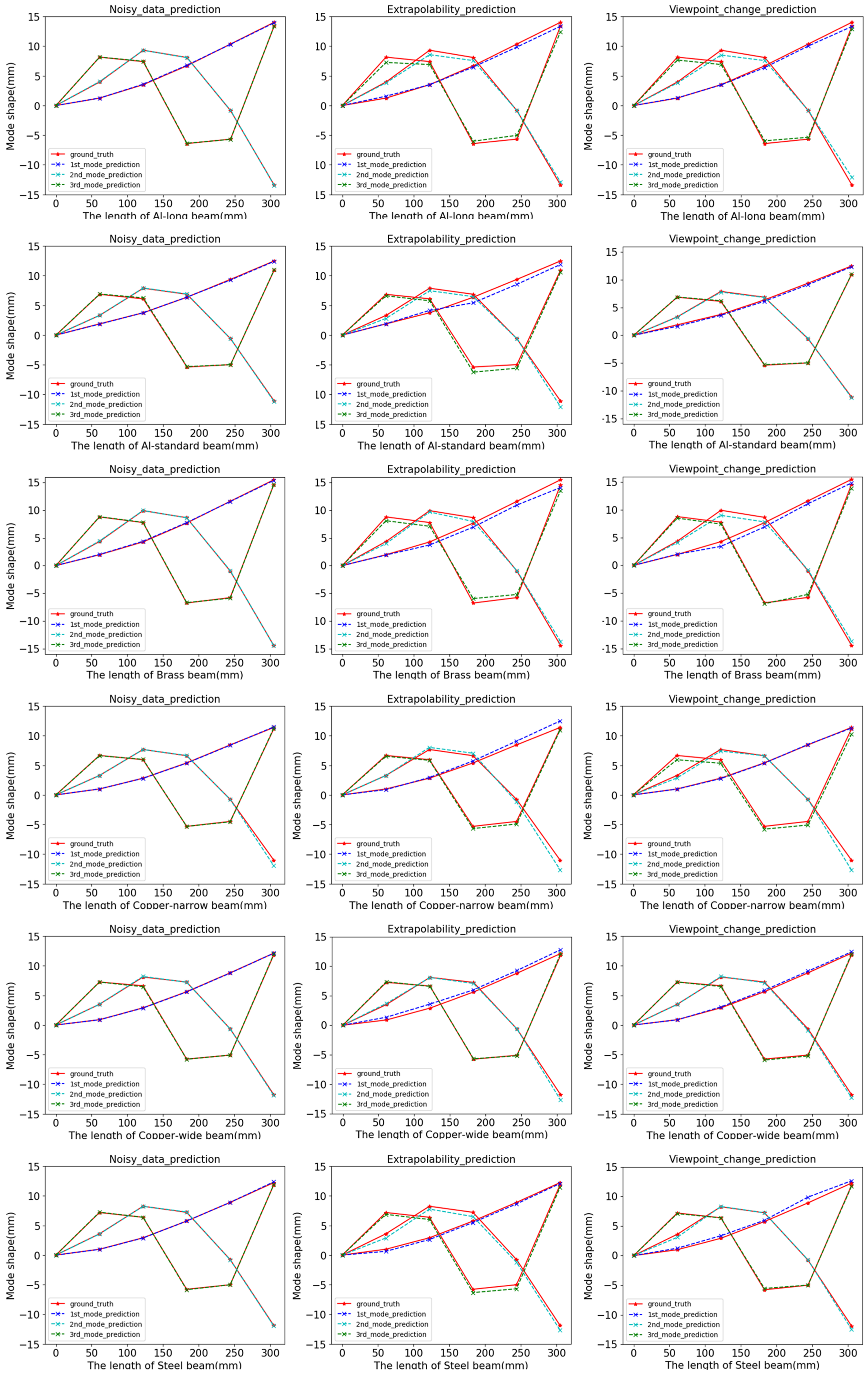

FIGURE 17 The ground truth and prediction value of six beam's mode shape

5.3 | Robustness on view-point change

Figure 16 depicts the MSE value calculation of the optimal CNN-LSTM model for the robustness for the view-point change task. Table 9 shows the mode shape prediction values of the CNN-LSTM model on robustness for view-point change in the video. The outlined CNN-LSTM model mimics the ground truth precisely. Table 10 states the three deep learning models' performance for viewpoint change task. Among all the beam specimens, our CNN-LSTM model gives the best results for the "Aluminum-standard" beam with an MSE value over the three mode shapes being 0.11. The model provides the poorest prediction for the "brass" sample. However, even in the "brass" sample, the model registers a good and acceptable MSE value of 0.34. The CNN-RNN and CNN-GRU achieve similar prediction results for all six beams, in which the average MSE values are 1.14 and 1.18, respectively.

The rightmost column of Figure 17 exhibits that the view-point change does not lead to results that are different from the ground truth.

6 | CONCLUSION AND FUTURE SCOPE

The paper outlines a CNN-LSTM deep learning model for a computer vision-based vibration estimation system that could be used to predict the mode shapes of various beam specimens. We utilized FEA based displacement values as the ground truth to compare the predictive performance of the CNN-LSTM model. The CNN-LSTM model is also compared with two architectures (1) CNN-RNN and (2) CNN-GRU. As is evident from the results, the performance of CNN-LSTM based computer vision model is comparable to the traditional procedures for the mode shape prediction task. It obtains better performance than two comparison models for all three different metrics since the LSTM unit can handle the information in memory for a more extended period than the conventional RNN unit. The outlined CNN-LSTM model showcases superior performance for robustness to noise and camera viewpoint change aspects. Likewise, on the extrapolability aspect (performance on unseen data) measurements, the CNN-LSTM model accomplishes palatable degrees of mode shape forecasting precision. The performance of the CNN-LSTM model gives support to the possible deployment of a non-contact computer vision-based approach for mode shape prediction problems.

The outlined work has concentrated on the excitation of various beams exposed to the time-varying load. The ground truth values of the first three mode shapes predicted by FEA are very close to the predictions made by our deep learning architecture. One future avenue for research in the SHM domain is to identify structural defects directly from the video recordings of the vibrating structure. Besides, future work can also focus on generating the pixel-wise displacement color map directly from the video stream of a vibrating structure. The produced displacement map could serve as a potential substitute for the FEA-based results of the vibration analysis.

ACKNOWLEDGEMENT

This work was partially supported by the Defense Advanced Research Projects Agency (DARPA), the Physics of Artificial Intelligence (PAI) program funding with agreement no. HR00111890037.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in https://github.com/ruoyuyang1991/ModeShape-dataset.

ORCID

Rahul Rai https://orcid.org/0000-0002-6478-4065

REFERENCES

1. Lynch JP, Loh KJ. A summary review of wireless sensors and sensor networks for structural health monitoring. *Shock Vib Dig*. 2006; 38(2):91-130.
2. Seo J, Hu JW, Lee J. Summary review of structural health monitoring applications for highway bridges. *J Perform Constr Facil*. 2016; 30(4):04015072.
3. OBrien EJ, Malekjafarian A. A mode shape-based damage detection approach using laser measurement from a vehicle crossing a simply supported bridge. *Struct Control Health Monit*. 2016;23(10):1273-1286.

4. Nguyen C-U, Huynh T-C, Kim J-T. Vibration-based damage detection in wind turbine towers using artificial neural networks. *Struct Monit Maint*. 2018;5(4):507.
5. Kim B, Min C, Kim H, Cho S, Oh J, Ha S-H, Yi J. Structural health monitoring with sensor data and cosine similarity for multi-damages. *Sensors*. 2019;19(14):3047.
6. Fan W, Qiao P. Vibration-based damage identification methods: a review and comparative study. *Struct Health Monit*. 2011;10(1):83-111.
7. Kim J-T, Ryu Y-S, Cho H-M, Stubbs N. Damage identification in beam-type structures: frequency-based method vs mode-shape-based method. *Eng Struct*. 2003;25(1):57-67.
8. Das S, Saha P, Patro SK. Vibration-based damage detection techniques used for health monitoring of structures: a review. *J Civ Struct Health Monit*. 2016;6(3):477-507.
9. Srinivasan C. *Structural Health Monitoring With Application to Offshore Structures*: World Scientific; 2019.
10. Bajrić A, Høgsberg J, Rüdinger F. Evaluation of damping estimates by automated operational modal analysis for offshore wind turbine tower vibrations. *Renew Energy*. 2018;116:153-163.
11. Sun M, Makki Alamdari M, Kalhori H. Automated operational modal analysis of a cable-stayed bridge. *J Bridg Eng*. 2017;22(12): 05017012.
12. Tarpø M, Nabuco B, Skafte A, Kristoffersen J, Vestermark J, Amador S, Brincker R. Operational modal analysis based prediction of actual stress in an offshore structural model. *Procedia Eng*. 2017;199:2262-2267.
13. Ni Y, Lu X, Lu W. Operational modal analysis of a high-rise multi-function building with dampers by a Bayesian approach. *Mech Syst Signal Process*. 2017;86:286-307.
14. Huang J, Zhou Z, Zhang L, Chen J, Ji C, Pham DT. Strain modal analysis of small and light pipes using distributed fibre Bragg grating sensors. *Sensors*. 2016;16(10):1583.
15. Zhang J, Maes K, De Roeck G, Reynders E, Papadimitriou C, Lombaert G. Optimal sensor placement for multi-setup modal analysis of structures. *J Sound Vib*. 2017;401:214-232.
16. Girolami A, Zonzini F, De Marchi L, Brunelli D, Benini L. Modal analysis of structures with low-cost embedded systems. In: 2018 IEEE International Symposium on Circuits and Systems (ISCAS) IEEE; 2018; Florence, Italy:1-4.
17. Xiong C, Lu H, Zhu J. Operational modal analysis of bridge structures with data from GNSS/accelerometer measurements. *Sensors*. 2017;17(3):436.
18. Theodosiou A, Lacraz A, Polis M, Kalli K, Tsangari M, Stassis A, Komodromos M. Modified fs-laser inscribed FBG array for rapid mode shape capture of free-free vibrating beams. *IEEE Photon Technol Lett*. 2016;28(14):1509-1512.
19. Yi T-H, Li H-N, Wang C-W. Multiaxial sensor placement optimization in structural health monitoring using distributed wolf algorithm. *Struct Control Health Monit*. 2016;23(4):719-734.
20. Yang J, Peng Z. Improved abc algorithm optimizing the bridge sensor placement. *Sensors*. 2018;18(7):2240.
21. Gomes GF, da Cunha SS, Alexandrino PSL, de Sousa BS, Ancelotti AC. Sensor placement optimization applied to laminated composite plates under vibration. *Struct Multidiscip Optim*. 2018;58(5):2099-2118.
22. Gomes GF, de Almeida FA, Alexandrino PSL, da Cunha SS, de Sousa BS, Ancelotti AC. A multiobjective sensor placement optimization for SHM systems considering Fisher information matrix and mode shape interpolation. *Engineering with Computers*. 2019; 35(2):519-535.
23. Huang C-H, Ma C-C. Experimental measurement of mode shapes and frequencies for vibration of plates by optical interferometry method. *J Vib Acoust*. 2000;123(2):276-280.
24. Linbo R, Geng T, Jing W, Haitao L, Hongfa H. Development of high-speed non-contact vibration measurement system. In: IEEE 2011 10th International Conference on Electronic Measurement & Instruments, Vol. 2 IEEE; 2011; Chengdu, China:244-247.
25. Schajer GS, Steinzig M. Sawblade vibration mode shape measurement using espi. *J Test Eval*. 2008;36(3):259-263.
26. Ehrhardt DA, Allen MS, Yang S, Beberniss TJ. Full-field linear and nonlinear measurements using continuous-scan laser doppler vibrometry and high speed three-dimensional digital image correlation. *Mech Syst Signal Process*. 2017;86:82-97.
27. Xu YF, Chen D-M, Zhu WD. Damage identification of beam structures using free response shapes obtained by use of a continuously scanning laser doppler vibrometer system. *Mech Syst Signal Process*. 2017;92:226-247.
28. Yang S, Allen MS. Output-only modal analysis using continuous-scan laser doppler vibrometry and application to a 20 kw wind turbine. *Mech Syst Signal Process*. 2012;31:228-245.
29. Xu Y, Brownjohn J, Kong D. A non-contact vision-based system for multipoint displacement monitoring in a cable-stayed footbridge. *Struct Control Health Monit*. 2018;25(5):e2155.
30. Molina-Viedma AJ, Felipe-Sesé L, López-Alba E, Díaz F. High frequency mode shapes characterisation using digital image correlation and phase-based motion magnification. *Mech Syst Signal Process*. 2018;102:245-261.
31. Patil K, Srivastava V, Baqersad J. A multi-view optical technique to obtain mode shapes of structures. *Measurement*. 2018;122:358-367.
32. Feng D, Feng MQ. Identification of structural stiffness and excitation forces in time domain using noncontact vision-based displacement measurement. *J Sound Vib*. 2017;406:15-28.
33. Javh J, Slavič J, Boltežar M. The subpixel resolution of optical-flow-based modal analysis. *Mech Syst Signal Process*. 2017;88:89-99.
34. Feng D, Feng MQ. Vision-based multipoint displacement measurement for structural health monitoring. *Struct Control Health Monit*. 2016;23(5):876-890.
35. Ye XW, Jin T, Yun CB. A review on deep learning-based structural health monitoring of civil infrastructures. *Smart Struct Syst*. 2019; 24(5):567-586.

36. Kohiyama M, Oka K, Yamashita T. Detection method of unlearned pattern using support vector machine in damage classification based on deep neural network. *Struct Control Health Monit*. 2020;27(8):e2552.
37. Tang Z, Chen Z, Bao Y, Li H. Convolutional neural network-based data anomaly detection method using multiple information for structural health monitoring. *Struct Control Health Monit*. 2019;26(1):e2296.
38. Khodabandehlou H, Pekcan G, Fadali MS. Vibration-based structural condition assessment using convolution neural networks. *Struct Control Health Monit*. 2019;26(2):e2308.
39. Xu Y, Wei S, Bao Y, Li H. Automatic seismic damage identification of reinforced concrete columns from images by a region-based deep convolutional neural network. *Struct Control Health Monit*. 2019;26(3):e2313.
40. Ni F, Zhang J, Chen Z. Pixel-level crack delineation in images with convolutional feature fusion. *Struct Control Health Monit*. 2019; 26(1):e2286.
41. Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T. Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2015:2625-2634.
42. Fan Y, Lu X, Li D, Liu Y. Video-based emotion recognition using CNN-RNN and C3D hybrid networks. In: Proceedings of the 18th acm international conference on multimodal interaction ACM; 2016; Tokyo, Japan:445-450.
43. Sainath TN, Vinyals O, Senior A, Sak H. Convolutional, long short-term memory, fully connected deep neural networks. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) IEEE; 2015; South Brisbane, QLD, Australia:4580-4584.
44. Alayba AM, Palade V, England M, Iqbal R. A combined CNN and LSTM model for arabic sentiment analysis. *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer; 2018:179-191.
45. Wang J, Yu L-C, Lai KR, Zhang X. Dimensional sentiment analysis using a regional CNN-LSTM model. In: *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*. Berlin, Germany; 2016:225-230.
46. Xu Z, Li S, Deng W. Learning temporal features using LSTM-CNN architecture for face anti-spoofing. *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. Kuala Lumpur, Malaysia: IEEE; 2015:141-145.
47. Yang R, Singh SK, Tavakkoli M, Amiri N, Yang Y, Karami MA, Rai R. CNN-LSTM deep learning architecture for computer vision-based modal frequency detection. *Mech Syst Signal Process*. 2020;144:106885.
48. Liang Y, Wu D, Liu G, Li Y, Gao C, Ma ZJ, Wu W. Big data-enabled multiscale serviceability analysis for aging bridges. *Digit Commun Netw*. 2016;2(3):97-107.
49. Zhang Z, Robinson D, Tepper J. Detecting hate speech on Twitter using a convolution-GRU based deep neural network. *European semantic web conference*. Cham: Springer; 2018:745-760.
50. Bradski G, Kaehler A. *Learning opencv: Computer vision with the opencv library*: “O'Reilly Media, Inc.”; 2008.
51. Yang R, Singh SK, Tavakkoli M, Amiri N, Karami MA, Rai R. Modeshape dataset. https://github.com/ruoyuyang1991/modeshape-dataset.git; 2019.
52. Claesen M, De Moor B. Hyperparameter search in machine learning. arXiv preprint arXiv:1502.02127; 2015.

How to cite this article: Yang R, Singh SK, Tavakkoli M, Amiri N, Karami MA, Rai R. Continuous video stream pixel sensor: A CNN-LSTM based deep learning approach for mode shape prediction. *Struct Control Health Monit*. 2022;29(3):e2892. doi:10.1002/stc.2892