A vision-based deep learning approach for non-contact vibration measurement using (2+1)D CNN and optical flow

Harold Harrison[1], **Mazlina Mamat**[2], **Farah Wong**[3], **Hoe Tung Yew**[4], **Racheal Lim**[5], **Wan Mimi Diyana Wan Zaki**[6]

[1, 2, 3]Applied Predictive Analytics Research Group, Faculty of Engineering, University of Malaysia Sabah, Sabah, Malaysia

[4]Electronic Engineering (Computer) Program, Faculty of Engineering, University of Malaysia Sabah, Sabah, Malaysia

[5]Sabah Electricity SDN. BHD., Wisma SESB, Jalan Tunku Abdul Rahman, 88673 Kota Kinabalu, Sabah, Malaysia

[6]Department of Electrical, Electronic and Systems Engineering, Faculty of Engineering and Built Environment, The National University of Malaysia, 43600 UKM Bangi, Selangor, Malaysia

[2]Corresponding author

E-mail: [1]*harold_harrison_dk24@iluv.ums.edu.my*, [2]*mazlina@ums.edu.my*, [3]*farrah@ums.edu.my*, [4]*htyew@ums.edu.my*, [5]*limracheal@yahoo.com*, [6]*wmdiyana@ukm.edu.my*

Received 16 April 2025; accepted 19 September 2025; published online 11 October 2025
DOI https://doi.org/10.21595/jve.2025.25002

Abstract. This paper introduces a proof-of-concept vision-based deep learning approach for vibration measurement, proposing a factorized (2+1)D Convolutional Neural Network (CNN) model to predict four vibration metrics: acceleration, velocity, displacement, and frequency, with a focus on rigid body motion. Unlike conventional neural network models that primarily focus on frequency prediction alone, this approach uniquely enables the simultaneous estimation of four critical vibration metrics, offering a comprehensive and cost-effective alternative to traditional contact-based sensors such as accelerometers. The framework relies on the visibility of a training fiducial marker, eliminates the need for calibration in controlled settings, enhancing scalability across specific environments. A curated dataset was generated using a controlled experimental setup comprising a single object in a lab-scale environment, augmented synthetically to enhance frequency diversity. An optical flow-based preprocessing algorithm synchronized motion features in recorded video inputs with measured vibration labels, improving measurement accuracy. The proposed model achieved an average Mean Absolute Percentage Error (MAPE) of 7.51 %, with acceleration predictions exhibiting the lowest error at 4.84 % and displacement the highest at 8.80 % across varying brightness levels and object-camera distances. Techniques such as Region of Interest (ROI) cropping and multi-section frame extraction were implemented to reduce computational complexity while further enhancing accuracy. These results highlight the framework's potential for non-invasive vibration analysis, though its generalizability is limited by the single-object dataset. Future work will expand the dataset, integrate multi-sensor inputs, explore marker-less tracking methods, and enable real-time deployment for predictive maintenance and structural health monitoring.

Keywords: vibration, non-contact, vision, deep learning.

1. Introduction

1.1. Background

The nature of the maintenance approach is moving from corrective to preventive and, in the future, predictive. Predictive maintenance detects trends and anticipates the problems before they arise [1]. This approach only performs equipment maintenance when necessary, thus saving maintenance costs and maximizing operation. Predictive maintenance uses condition-based monitoring, machine learning, and analytics techniques to predict upcoming machine or asset failures. Condition-based monitoring (CBM) requires high-performance sensors, including accelerometers, pressure sensors, MEMS microphones, etc. [2], [3]. However, the most functional sensors are the accelerometers for detecting vibration, an important starting parameter for predictive maintenance [4]. In the context of vibration detection, conventional piezoelectric accelerometers, often regarded as the gold standard, are less suitable for wireless predictive maintenance systems due to limitations in size, lack of integrated features, and high-power consumption [5]. Additionally, the installation of these sensors may be difficult and expensive [6-8]. To overcome these limitations, non-contact sensing techniques have been developed and can be broadly categorized into laser-based, vision-based, and deep learning-driven approaches. A widely used laser-based method is the laser Doppler vibrometer (LDV), which offers high precision but suffers from drawbacks related to cost, bulkiness, and limited accessibility [9]. Moreover, since a laser is a concentrated beam of energy, it can raise the surface temperature of the object under measurement, making it potentially invasive and capable of altering the object's natural vibration frequency [10, 11]. As an alternative, vision-based sensing methods have gained attention due to their low cost, ease of deployment, and non-invasiveness. These include techniques such as digital image correlation (DIC), optical flow analysis, and deep learning-based models that infer structural vibration responses directly from video data. Among these, vision-based neural networks are emerging as a promising solution for modal analysis and multi-parameter vibration estimation.

1.2. Existing methods of non-contact vibration measurement

1.2.1. Laser-based measurement

Laser-based measurements, particularly laser triangulation and Laser Doppler Vibrometry (LDV), have been widely studied. Laser triangulation operates by projecting a laser beam onto the vibrating surface and capturing the reflected spot using a camera. The displacement of the laser spot within the camera's field of view is analyzed geometrically to estimate surface vibrations. This method is particularly effective for high-precision, small-amplitude vibration measurements, such as those required in turbine blade monitoring [12]. However, its measurement range is constrained by the optical system’s depth of field, and it may struggle with large structures, highly curved surfaces, or large displacements.

LDV, in contrast, relies on detecting the Doppler shift of laser light reflected from a moving surface to measure surface velocity directly. LDV offers excellent frequency response and high accuracy across a broad dynamic range. Nonetheless, system performance can be affected by environmental factors such as extreme temperatures, which may degrade the stability and alignment of optical components. LDV systems also depend on the optical reflectivity of the target surface; poorly reflective or highly diffusive materials often require surface treatment, such as applying reflective tape or coatings, which can be time-consuming and infeasible in certain applications [13]. Although LDV systems can capture both low and high-frequency vibrations, their effectiveness at very low amplitudes may be limited by signal-to-noise ratio constraints. Furthermore, the high cost of acquisition and maintenance can be prohibitive for small laboratories or budget-constrained projects.

1.2.2. Vision-based sensing methods

High-speed imaging techniques have emerged as valuable tools for capturing dynamic responses in structural vibration analysis. By recording rapid sequences of images, high-speed cameras enable the frame-by-frame tracking of object motion, facilitating precise estimation of displacement and vibrational frequency. This approach provides high temporal resolution and visual interpretability, making it particularly advantageous in experimental mechanics. However, the method necessitates the use of sophisticated and costly equipment, including high-frame-rate cameras and adequate lighting systems. For instance, prior studies employing industrial-grade imaging systems operating at 300 frames per second have demonstrated reliable measurement of vibrational motion under controlled conditions [14]. While this frame rate is sufficient for low-to-mid frequency phenomena, higher-speed dynamics require significantly greater temporal resolution.

Digital Image Correlation (DIC) represents a non-contact, full-field optical measurement technique that is widely adopted in materials science, structural health monitoring, and civil engineering. The method operates by analyzing the displacement of a random speckle pattern applied to a surface and comparing sequential digital images acquired during dynamic loading. Through cross-correlation algorithms and sub-pixel interpolation techniques, DIC provides accurate spatial mapping of displacement and deformation fields [15]. High-speed DIC (HS-DIC) has proven effective for modal analysis and vibration characterization, particularly in applications involving large surfaces or complex geometries. The method offers distinct advantages over point-based techniques, such as LDV, by capturing distributed vibration modes with full-field resolution [16].

Nonetheless, HS-DIC presents practical challenges that have limited its widespread industrial adoption. These include the complexity of experimental setup, the requirement for high-quality stochastic patterns, sensitivity to environmental lighting, and the necessity for robust image acquisition and processing pipelines [16]. Additionally, while DIC provides adequate displacement sensitivity for many engineering applications, its resolution is typically lower than that of LDV systems [17, 18].

The accuracy and reliability of DIC measurements are heavily influenced by camera resolution, frame rate, optical quality, and the performance of the correlation algorithm. Sub-pixel interpolation enhances spatial resolution but introduces sensitivity to subset size, image noise, and systematic errors from lens distortion or misalignment [19]. Despite these limitations, DIC remains a powerful tool for full-field vibration analysis, particularly in environments where LDV performance may be degraded by speckle noise, non-reflective surfaces, or restricted optical access.

Optical flow refers to the apparent motion of features within an image sequence resulting from relative motion between the observer and the observed scene. In the context of vibration measurement, optical flow algorithms are applied to successive frames of high-speed video data to estimate pixel-level displacements and velocities of vibrating structures. By analyzing the temporal motion vectors, it is possible to reconstruct vibrational signals and extract frequency content through spectral analysis techniques such as the Fast Fourier Transform (FFT). Modern optical flow techniques can achieve sub-pixel displacement resolution, making them particularly suitable for applications in structural health monitoring, where early detection of micro-scale vibrations is critical for identifying potential structural anomalies [20]. Furthermore, phase-based optical flow methods have emerged as a significant advancement over traditional intensity-based approaches. These methods operate in the frequency domain, leveraging local phase information to detect subtle motions more robustly. Derivative-enhanced phase-based optical flow (PBOF) algorithms have demonstrated superior accuracy in visual vibration measurement tasks [21-23].

Despite these advantages, the application of optical flow to vibration analysis presents certain limitations. The measurable displacement range is inherently constrained by the algorithm’s sensitivity and the frame-to-frame motion resolution. Additionally, the maximum detectable velocity is bounded by the frame rate of the imaging system, as large inter-frame displacements may violate the assumptions of optical flow estimation [24]. Displacement scaling must also be calibrated with respect to the camera-to-object distance, either through manual or automated calibration procedures. A fundamental limitation of most optical flow methods lies in the brightness constancy assumption, which assumes that pixel intensities remain constant across frames. This assumption becomes problematic in dynamic lighting conditions or on reflective and deformable surfaces, potentially reducing the accuracy of motion estimation. While Phase-Based

Optical Flow (PBOF) mitigates this issue, it incurs significant computational costs, requiring Fourier transforms for each frame [25]. Addressing these challenges requires either robust preprocessing, adaptive models, or integration with complementary sensing techniques to ensure reliability in uncontrolled environments.

1.2.3. Neural network model vibration measurement

Artificial intelligence (AI) techniques, particularly those utilizing unsupervised learning paradigms such as autoencoders and clustering algorithms, have been increasingly applied to anomaly detection in vibration data. These approaches are advantageous in scenarios where labeled datasets are scarce or costly to obtain, as they can model normal operational behavior and identify deviations that may indicate mechanical faults [26]. Unlike traditional approaches that aim to quantify specific vibration parameters, unsupervised methods are primarily used to detect abnormalities in signal patterns, serving as early indicators of potential failure.

In parallel, supervised learning models have been extensively employed to classify specific fault types based on labeled vibration signals acquired through contact-based sensors [27, 28]. Furthermore, supervised regression models have been explored for predicting vibration acceleration values by learning from historical sensor data, thereby facilitating data-driven condition monitoring [29]. These methods typically require curated datasets but offer higher precision in both classification and regression tasks when sufficient labeled data are available.

Recent advancements have extended supervised learning into vision-based domains, where deep convolutional neural networks (CNNs) are trained to estimate vibration frequency from video sequences. One notable study trained a CNN on synthetically generated signals and demonstrated its ability to infer vibration frequencies by analyzing pixel-level brightness variations over time [30]. The network achieved acceptable frequency prediction in the range of 1-30 Hz using video recorded at 100 frames per second, consistent with Nyquist sampling constraints [31]. However, the study was limited to frequency estimation without addressing other key parameters such as displacement, velocity, or acceleration. It also did not fully account for noise, varying illumination conditions, or the selection of optimal pixels for signal extraction. In another approach, a hybrid CNN-LSTM model was proposed to predict modal frequencies by treating each pixel in the video frame as a virtual sensor, capturing both spatial and temporal characteristics of structural vibrations [32].

As for deep learning categories, many of the solutions to vibration measurement lean toward identifying faulty machinery components and are less focused on giving values of vibration measurements. Most deep-learning studies in this domain remain focused on machine fault classification or frequency estimation from sensor data. There is a distinct gap in vision-based deep learning frameworks capable of directly regressing comprehensive vibration parameters: specifically, acceleration, velocity, and displacement from raw video inputs.

1.3. Paper contribution

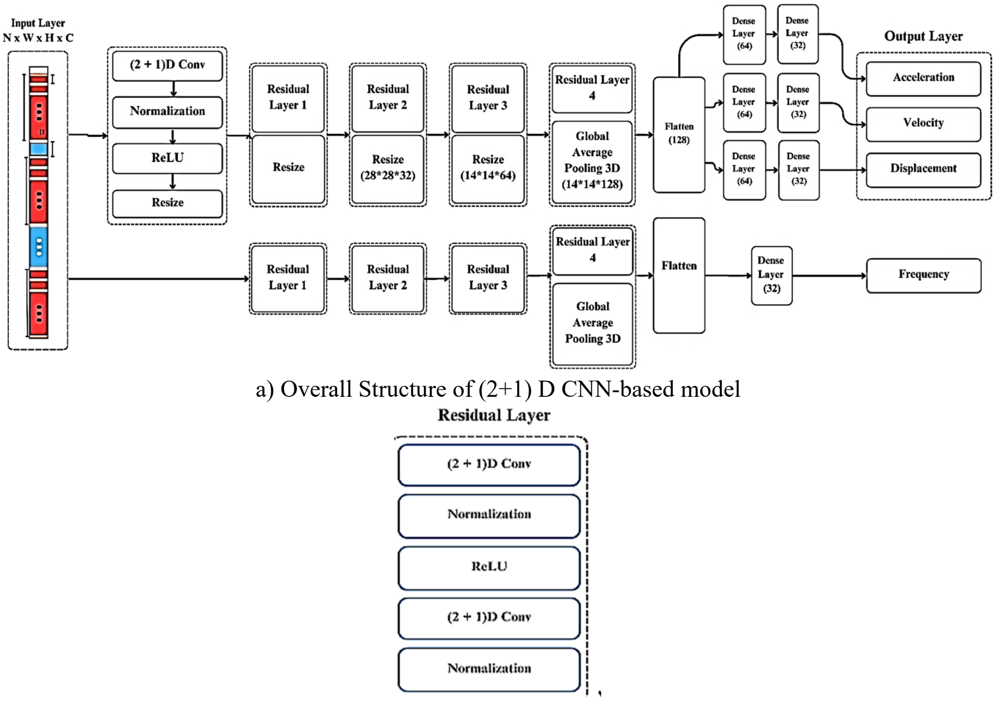
Despite the potential of 3D CNNs for vibration measurement, no research has explored their application or developed models based on this framework in this domain. Existing studies have primarily relied on CNN-LSTM models, which focus predominantly on shape frequency prediction and lack a comprehensive analysis of other vibration metrics. To address this gap, this study proposes leveraging recent advancements in 3D CNNs, specifically the factorization of kernels into the (2+1)D CNN architecture, to enhance vibration measurement analysis.

This paper presents a novel non-contact vibration measurement framework that uniquely combines a (2+1)D CNN with optical flow preprocessing to simultaneously predict acceleration, velocity, displacement, and frequency from video inputs. Unlike prior vision-based methods that address only one or two metrics, our approach delivers a holistic solution by regressing all four key vibration parameters, marking a significant departure from conventional techniques. The integration of optical flow addresses critical data desynchronization issues, enabling precise temporal alignment between video frames and vibration signals. The model also aims to adapt in varying light levels and camera-object distances without the calibration required by pure image processing methods like DIC or optical flow, ensuring robustness across diverse environments.

2. Methodology

2.1. Four-output regression (2+1) dimension CNN

Predicting acceleration, velocity, displacement, and frequency from sequences of images requires effective modeling of temporal information. Traditional two-dimensional (2D) CNNs are limited in this regard, as they process frames independently and cannot capture temporal dynamics. Three-dimensional (3D) CNNs address this limitation by extracting spatiotemporal features across consecutive frames; however, they introduce substantial computational overhead. To balance temporal modeling capability with computational efficiency, a 3D CNN can be factorized into a (2+1)D CNN, which separates spatial and temporal convolutions while preserving temporal awareness [33]. This architecture supports the use of 3D CNNs over alternatives such as CNN–LSTM models, which, although previously applied to modal frequency prediction, were restricted to shape frequency and did not address other key vibration parameters.



a) Overall Structure of (2+1) D CNN-based model

b) Contents of the residual layer

Fig. 1. Overall Structure of the proposed model and its residual layer content

Fig. 1 illustrates the architecture of the proposed four-output regression model based on (2+1)D CNN. The model comprises several distinct processing stages designed to handle spatiotemporal vibration data analysis. The input layer accepts three-dimensional data, where the first two dimensions represent spatial frame information (height × width) and the third dimension corresponds to temporal information (number of sequential frames). This input undergoes parallel processing through two distinct convolutional operations: a 2D spatial convolution analyzing frame content and a 1D temporal convolution examining pixel evolution across frames. Following the input layer, the (2+1)D CNN layer performs initial feature extraction, subsequently normalized and down-sampled through dimensionality reduction.

The architecture then incorporates multiple residual blocks, each performing progressive downscaling. These residual connections serve two critical functions: mitigating model degradation during deep network training and enhancing discriminative feature extraction. The hierarchical downscaling process enables the model to selectively focus on relevant frame regions while suppressing extraneous information, including background noise and lighting artifacts. The final stages employ 3D pooling operations followed by feature vector flattening, transforming the spatiotemporal data into a compact representation. This distilled feature set feeds into four specialized hidden layers, each optimized for predicting one of the four target vibration metrics.

The (2+1)D CNN architecture was selected for its balance of computational efficiency and temporal modeling capability.

2.2. Equipment and experiment setup

The experiment setup is shown in Figs. 2 and 3, which consists of a vibration motor, a vibration meter, an accelerometer sensor, a Raspberry Pi 5 with a camera, an illumination source, and a computer. At the core of this setup is a vibration motor, which generates controlled oscillatory motion. The vibrations affect a cube to which the motor is attached. To quantify these vibrations, a vibration meter and an accelerometer (MPU6050) are placed directly in contact with the cube to get the most accurate reading of acceleration, velocity, and displacements. The accelerometer continuously records acceleration data over time, which is later processed to compute vibration frequency.

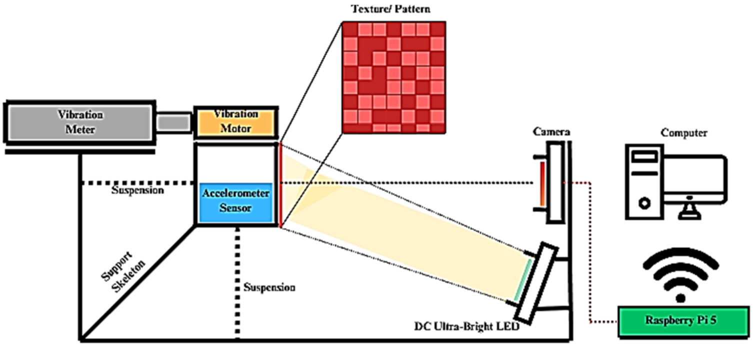


Fig. 2. Experiment setup diagram

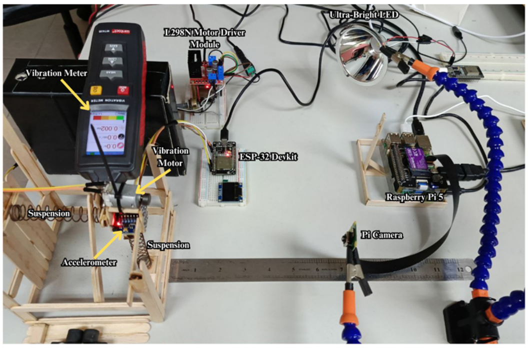
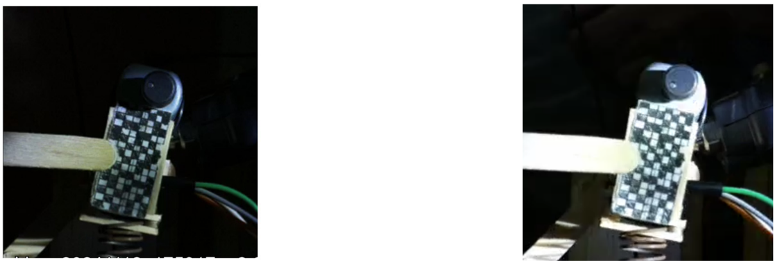


Fig. 3. Actual setup. Image credit: Harold Harrison, UMS Faculty of Engineering Laboratory, 2024

This setup employs two ESP-32 microcontrollers to ensure reliable operation. The first microcontroller regulates the vibration motor to control the magnitude of vibration velocities, while the second records acceleration data from the accelerometer. Using separate ESP-32 units prevents interference between motor control and sensor sampling, which could otherwise compromise measurement accuracy. Although parallel programming on a single ESP-32could theoretically achieve similar functionality, preliminary testing showed that multi-threading leads to excessive heat generation causing system instability. Furthermore, this study found that combining high-frequency sensor sampling with motor control on a single microcontroller resulted in frequent system crashes.

An essential component of the experimental setup involves applying a textured pattern to the surface of the cube to enhance motion tracking accuracy by improving shift and displacement detection. A Raspberry Pi 5 camera, positioned at a distance from 10 to 50 cm perpendicular to the cube, captures video footage at high frame rates. The video stream is transmitted in real time to a dedicated processing computer via a Flask-based server system, enabling immediate preprocessing and model training.

To maintain image clarity under high shutter speeds, an external DC-powered ultra-bright LED light source provides consistent illumination. Data collection is conducted under two distinct environmental conditions: (1) a semi-controlled laboratory setting with variable ambient lighting (including natural light from windows and artificial ceiling lighting) and (2) an open space to simulate real-world scenarios. This approach ensures dataset diversity in illumination levels, further augmented by varying LED light levels. Representative video frames illustrating different lux configurations are provided in Fig. 4, demonstrating the range of illumination conditions incorporated into the dataset.



a) b)

Fig. 4. Single frame at a) 50 lux, b) 300 lux.
Image credit: Harold Harrison, UMS Faculty of Engineering Laboratory, 2024

Visualizing vibrational motion requires the analysis of four fundamental kinematic parameters: displacement, velocity, acceleration, and frequency. Displacement is determined by tracking the positional coordinates of a reference point over time, with temporal resolution provided by synchronized timestamps. Velocity and acceleration are subsequently derived through first- and second-order temporal differentiation of the displacement data, respectively. Frequency characterization, representing the system's predominant oscillation rate, is typically obtained via spectral analysis of periodic motion. This is most commonly achieved through computational methods such as the Fast Fourier Transform (FFT) or related signal processing techniques that decompose the time-domain signal into its constituent frequency components.

Accurate frequency measurement requires adherence to the Nyquist sampling theorem, which specifies that the sampling rate must be at least twice the highest frequency component of interest. Given that typical machinery operates below 60 Hz [34], a minimum sampling rate of 120 Hz is required for reliable measurement. In the context of optical measurement, the sampling rate corresponds to the camera's frame rate, necessitating a capture capability of at least 120 frames per second (fps). This study employs a Raspberry Pi 5 equipped with a high-speed camera module capable of exceeding 120 fps, thereby satisfying the Nyquist criterion for the target frequency range. Theoretically, this configuration enables accurate vibration measurement across the 1-60 Hz spectrum, encompassing the operational frequencies of most mechanical systems.

2.3. Acquisition of videos, vibration data and preprocessing

The vibration measurement and video recording methodology are illustrated in Fig. 5. A calibrated vibration meter is employed to quantify vibrations induced by the motor, with the sensor positioned atop the cube in direct structural contact. This configuration ensures optimal sensor coupling while accounting for the sensor's mass contribution to the overall system load, thereby improving measurement consistency. The vibration meter provides direct readings of acceleration, velocity, displacement, and frequency range.

For precise frequency determination, a triaxial accelerometer (MPU6050) with a 300 Hz sampling capability supplements the vibration meter measurements. During synchronous video recording, the acquired acceleration data is logged for subsequent spectral analysis via Fast Fourier Transform (FFT) (Fig. 6). To validate measurement integrity, FFT-derived frequencies are cross-verified against the vibration meter's indicated range; any discrepancies result in data exclusion from further analysis.

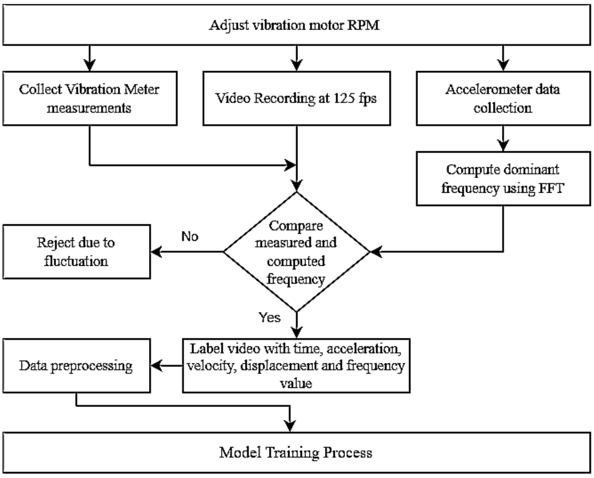

Adjust vibration motor RPM
Collect Vibration Meter measurements
Video Recording at 125 fps
Accelerometer data collection
Compute dominant frequency using FFT
Reject due to fluctuation
No
Compare measured and computed frequency
Yes
Data preprocessing
Label video with time, acceleration, velocity, displacement and frequency value
Model Training Process

Fig. 5. Data collection flow-chart

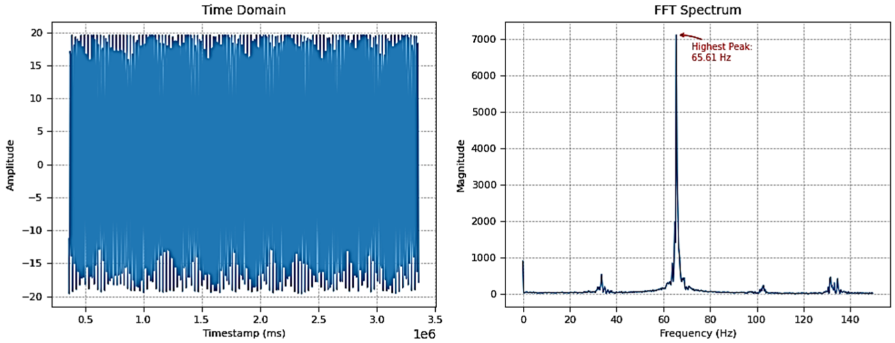


Fig. 6. Time-domain signal and corresponding FFT spectrum of MPU6050 acceleration data

As the primary goal is to establish proof-of-concept feasibility, the MPU6050's programmable ranges (±2 g to ±16 g) and 16-bit resolution provide adequate performance to capture general vibration trends in controlled settings, offering sufficient data to demonstrate correlations between video inputs and vibration outputs during the initial development phase. Furthermore, its use establishes a scalable baseline, paving the way for future enhancements by identifying model strengths and weaknesses that can be addressed with more precise sensors in subsequent work. For future research, it is recommended to incorporate high-precision equipment, such as industrial-grade accelerometers, to improve accuracy and robustness as the model progresses toward real-world deployment.

The Raspberry Pi captures video footage at a frame rate of 125 fps. Certain settings were adjusted to enable high-speed frame capturing. One key modification was reducing the resolution to 640×640 pixels to minimize the GPU workload. The shutter speed was set to 1/300 seconds, and the Libav codec was selected, as other codecs impose a maximum recording frame rate of less than 120 fps. Additionally, the default denoise feature of the libcamera-vid library, which is commonly used with the Pi camera, was disabled to conserve GPU resources during video recording. A frame processing function was developed to process all recorded video data efficiently. This function accepts several parameters, including the frame resolution, the total number of frames, and the spacing between extracted frames. During development, the resolution and frame count were carefully tuned and optimized to enhance performance. The frame processing function begins by scanning video files that adhere to a specific naming convention: a{A}_v{V}_d{D}_f{F}.mp4', where 'A' represents acceleration, 'V' represents velocity, 'D' represents displacement, and 'F' represents frequency measurements.

To address the limited size of the dataset, particularly concerning the frequency spectrum, data augmentation was performed by rendering the same collected videos at varying playback speeds (either faster or slower). This approach expanded the frequency range of the dataset to span 5 to 80 Hz. The resulting videos are all rendered at a constant 300 fps for the sole purpose of testing if the frame speed does affect the range of frequency that the model can predict. Additionally, to enhance diversity in illumination conditions, videos were captured using different shutter speeds, thereby increasing variability in lighting levels across the dataset.

Each video filename encapsulates all relevant vibration measurement labels, ensuring clarity in dataset organization. Once identified, frames are extracted at a specified depth (i.e., the number of frames input into the model). The function also supports multi-section frame extraction, allowing multiple segments of frames to be retrieved from a single video, maximizing the utilization of the dataset. For example, in a 4-second video recorded at 125 frames per second (fps), a total of 500 frames is available. The multi-sectioning method enhances dataset efficiency by enabling multi-sampling of frames, reducing frame wastage. Fig. 7 illustrates this multi-sectioning approach in detail.

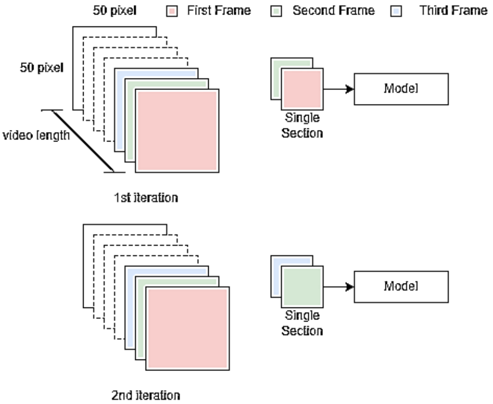


Fig. 7. Multi-section frame sampling of a depth of two in a single video

Once the frames and labels are organized, the dataset is split into a training set (80 %) and a validation set (20 %). Within the validation dataset, a sub-segment is designated as the testing dataset, which is used to evaluate and showcase the model's performance. Initial data collection consists of 100 4-second videos. After going through the optical flow validation algorithm, around 2000 videos were created, consisting of five frames in each video while maintaining the same time interval between each frame.

2.4. Performance metric

For model evaluation, the Mean Absolute Percentage Error (MAPE) metric is employed alongside the Root Mean Square Error (RMSE) and the Coefficient of Determination (R^2). The rationale for choosing MAPE is that it enables the detection of individual measurements that may underperform or outperform expectations, providing a more interpretable measure of error in percentage terms. Since the dataset contains measurements at varying scales, MAPE offers a more standardized error assessment [35], making it a suitable loss function for model training. RMSE is included to measure the average magnitude of the errors in the predicted values, giving higher weight to larger errors, which is useful for understanding the spread of prediction errors. R^2 is utilized to assess the proportion of variance in the dependent variable that is predictable from the independent variables, providing insight into the goodness of fit of the model.

To be considered reliable, the MAPE must be below 30 %, as values beyond this threshold suggest that the model is engaging in random guessing rather than effectively generalizing the dataset. The performance is considered acceptable if MAPE falls under 30 %, good within the range of 10 % to 20 %, and excellent performance if it falls below 10 %. For RMSE, a lower value indicates better performance, with the threshold for acceptability depending on the scale of the data. An R^2 value closer to 1 indicates a better fit, with values above 0.7 generally considered acceptable, above 0.9 indicating good performance, and values near 1 reflecting excellent performance.

3. Results and discussions

3.1. Initial model

The input to the initial model is the consecutive frames without any pre-processing. Initially, the training starts with two depth frames, which are gradually increased to 60 depth frames to determine the optimal performance. Several input resolutions are also tested to identify the most effective configuration. The optimization process involves tuning the (2+1)D CNN hyperparameters, including convolutional layers, pooling layers, fully connected layers, frame resolution, and the number of input frames. Table 1 shows the performance of the initial model, and the results indicate poor performance.

Table 1. Prediction performance of the initial model

Metrics	Acceleration (m/s^2)	Velocity (mm/s)	Displacement (mm)	Frequency (Hz)
MAPE (%)	42.99 %	43.08 %	44.76 %	57.81 %
RMSE	5.3308	48.2907	0.4269	14.4496
R^2	–0.1135	–24.3154	–19.4189	–0.3982

Upon further analysis, the primary cause of the observed performance degradation was identified as a temporal misalignment between vibration measurements and video frame motion. This discrepancy stemmed from inherent asynchrony in the data acquisition systems: the camera and accelerometer operated on distinct internal timing mechanisms during data collection. As depicted in Fig. 8, the optical flow-derived motion from the video frames deviates from the accelerometer-recorded vibrations, with a pronounced offset in their initial timestamps. This desynchronization led to erroneous correlations between the visual motion in frames and their corresponding vibration values, compromising the accuracy of subsequent analyses.

An initial corrective measure involved applying a fixed delay to the video recording to compensate for the temporal desynchronization. However, subsequent analysis revealed that the necessary delay exhibited significant variability across recordings, rendering this approach inconsistent and unreliable. Consequently, a more robust and adaptive synchronization methodology was deemed necessary to ensure precise temporal alignment.

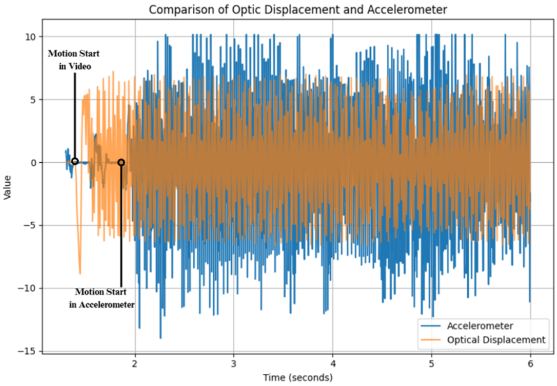


Fig. 8. Vibration measurement labeling error

3.2. Improved model

To improve prediction accuracy, a Region of Interest (ROI) is defined to isolate the vibrating object and suppress noise from background motion, as illustrated in Fig. 9. This spatial filtering ensures that only the most relevant pixel data contributes to the feature extraction process.

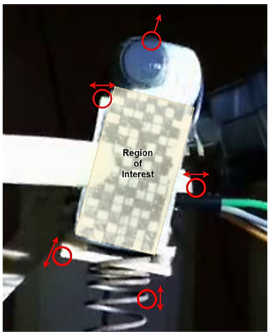


Fig. 9. Motions outside of ROI

To resolve the desynchronization issue, an image preprocessing algorithm based on optical flow was implemented. This technique analyzes sequential frames in the video to detect and track pixel motion, effectively aligning video data with vibration measurements. The algorithm selects pixels located on textured regions or edges, where motion estimation is more reliable. Displacement is computed using the Euclidean Distance Formula by tracking pixel coordinates across frames. Velocity is then calculated by dividing displacement by the time interval, which remains constant at 1/125 seconds, corresponding to the 125-fps sampling rate. Acceleration is computed by evaluating the change in velocity across three consecutive frames. For frequency analysis, at least five frames are required to perform a Fast Fourier Transform (FFT), enabling the extraction of dominant vibration frequencies. This preprocessing strategy significantly improves temporal alignment between video and vibration data, ensuring more accurate and efficient feature extraction. A schematic representation of the optical flow algorithm is provided in Fig. 10.

To verify the effectiveness of this approach, the motion-derived values obtained through optical flow were compared against measurements from a calibrated vibration meter. The integration of optical flow notably reduced the number of input frames needed for accurate vibration analysis. Whereas the initial model required 30 or more frames to achieve reliable predictions, the optical flow-enhanced pipeline was able to compute displacement and velocity using only two frames, and acceleration and frequency using five frames. This optimization reduced the minimum frame requirement from 30 to 15, significantly enhancing computational efficiency without compromising prediction accuracy.

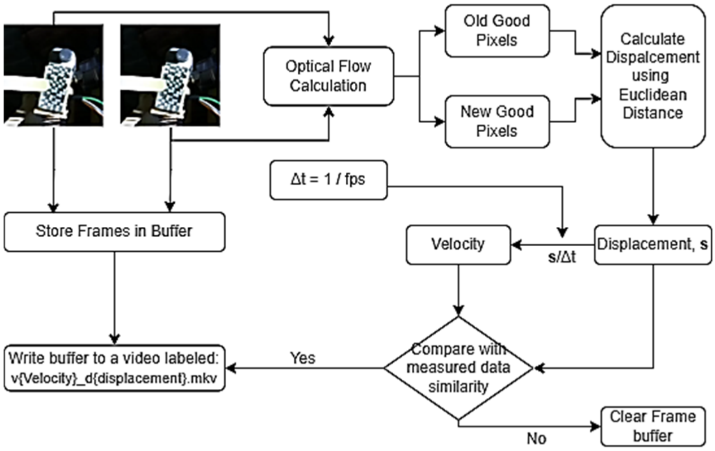

Optical Flow Calculation
Old Good Pixels
New Good Pixels
Calculate Dispalcement using Euclidean Distance
Δt = 1 / fps
Store Frames in Buffer
Velocity
s/Δt
Displacement, s
Write buffer to a video labeled: v{Velocity}_d{displacement}.mkv
Yes
Compare with measured data similarity
No
Clear Frame buffer

Fig. 10. Optical flow validation flowchart

Table 2 presents the performance evaluation of the improved model, demonstrating significant enhancements across all metrics when tested on the same dataset. Among the vibration metrics, velocity exhibited the best performance, achieving a MAPE of 3.46 %, followed closely by displacement at 3.77 %. Acceleration ranked third with a MAPE of 4.53 %, while frequency recorded the highest error, with an MAPE of 11.76 %, indicating room for improvement. The RMSE values, which measure the average magnitude of errors, were 0.6979 for acceleration, 4.3467 for velocity, 0.0410 for displacement, and 4.5361 for frequency, highlighting the spread of prediction errors. The R^2 values, indicating the proportion of variance explained by the model, were 0.9809 for acceleration, 0.7949 for velocity, 0.7960 for displacement, and 0.8551 for frequency, reflecting the goodness of fit across these metrics.

One possible explanation for the performance disparity in frequency estimation is the lack of dataset variability. The collected frequency data is constrained to the natural frequency of the object, limiting the model's exposure to diverse vibration patterns. This study focused on a single type of object with a constant mass and material composition, reducing the dataset's generalizability. A thorough examination of the dataset revealed only four dominant vibration frequencies: 15.25 Hz, 30.50 Hz, 34.86 Hz, and 45.75 Hz. These frequencies align with values reported in studies on kinematic structure performance [34], which highlight common natural frequencies in machine vibrations.

The model's performance improved significantly following dataset augmentation with an expanded range of vibration frequencies and varied light illumination levels, as evidenced by the test results presented in Table 3. These enhancements, combined with optimizations across multiple model pipelines, yielded a 35.0 % reduction in MAPE from 11.76 % to 7.64 %, demonstrating improved accuracy in frequency metric prediction. The RMSE decreased substantially by 62.0 %, from 4.5361 to 1.7334, indicating markedly reduced prediction errors. Furthermore, the R^2 increased by 16.4 percentage points to 0.9954, showing near-perfect variance explanation. These improvements suggest that incorporating broader frequency spectra and illumination variability, along with pipeline optimization, effectively enhances the model's generalization capability. While the current MAPE of 7.64 % may be acceptable for many applications, further refinement through techniques like weighted loss functions could potentially yield additional gains in prediction accuracy.

Fig. 11 presents the R^2 values for each measurement metric, evaluated using a dataset incorporating variable lighting conditions and an extended frequency range. The model demonstrates robust predictive capability, explaining over 80 % of the observed variance despite the presence of outliers.

Table 2. Model performance and predictions with optical flow validation and ROI cropping

Metrics	Acceleration (m/s^2)	Velocity (mm/s)	Displacement (mm)	Frequency (Hz)
MAPE (%)	4.529	3.463	3.768	11.76
RMSE	0.6979	4.3467	0.0410	4.5361
R^2	0.9809	0.7949	0.7960	0.8551

Table 3. Model performance on the dataset with light and frequency variants

Metrics	Acceleration (m/s^2)	Velocity (mm/s)	Displacement (mm)	Frequency (Hz)
MAPE (%)	4.8441	8.7617	8.8097	7.6425
RMSE	0.5863	6.3810	0.0623	1.7334
R^2	0.9658	0.8916	0.9004	0.9954

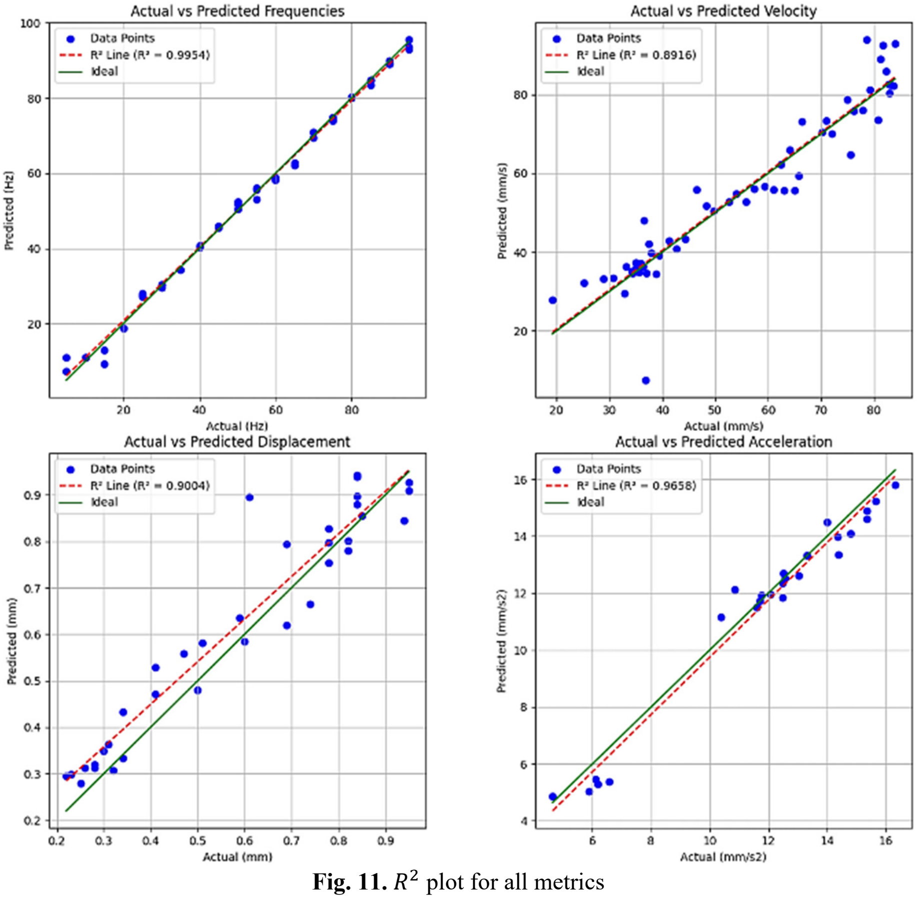


Fig. 11. R^2 plot for all metrics

Detailed analysis of these outliers reveals systematic patterns in their distribution. The predominant cluster occurs at shutter speeds exceeding 700, where diminished light availability at the Raspberry Pi camera sensor significantly compromises feature extraction. Furthermore, the velocity metric exhibits a progressive increase in outlier frequency toward the upper range of measured values, attributable to motion blur artifacts when subject movement exceeds the camera's temporal resolution. This relationship between motion velocity and outlier prevalence is similarly reflected in the displacement metric, where higher displacement measurements demonstrate greater deviation from model predictions, consistent with the expected limitations of optical measurement under rapid motion conditions.

3.3. Effect of input size, network depth, distance to camera, and fiducial markers

Table 4 presents the MAPE values across different input resolutions and network depths, both with and without ROI cropping. The findings indicate that higher-resolution inputs generally lead to lower MAPE values, suggesting that increasing the input size enhances the model’s ability to extract meaningful vibration features. For instance, a resolution of 50×50 pixels consistently achieves a lower MAPE than 32×32 pixels. Increasing the number of layers improves performance, with deeper networks (four layers) consistently yielding lower MAPE values compared to shallower networks (two or three layers). This suggests that deeper architectures can better extract complex motion patterns from vibration data.

Table 4. Average MAPE across different model settings

Input size, pixels	Number of layers	Initial model	Optical flow validation	Optical flow validation + ROI Cropping
32 * 32	4	55.16	23.59	9.74
50 * 50	4	56.37	18.31	7.51
32 * 32	3	50.70	21.04	12.13
50 * 50	3	51.64	19.91	8.72
32 * 32	2	60.08	24.01	16.89
50 * 50	2	58.56	23.85	16.01

A critical observation is a dramatic reduction in error rates with ROI cropping, particularly for lower-resolution inputs. For example, at 50×50 pixels with four layers, MAPE drops from 56.37 % to just 5.93 % after applying optical flow validation and ROI cropping. This highlights the effectiveness of focusing on the core vibrating object, eliminating irrelevant background information, and improving model precision [36]. This analysis suggests that higher input resolutions, deeper networks, and ROI cropping significantly enhance performance. ROI cropping plays a crucial role in reducing error rates by isolating relevant motion areas.

As demonstrated in Table 5, the model exhibited significantly degraded performance when the fiducial marker was occluded, with prediction accuracy declining proportionally with increasing camera-to-object distance. This finding strongly suggests that the model has learned to utilize the marker as a key visual feature for motion estimation, effectively employing it to dynamically adjust the mm/pixel conversion parameters across varying distances. However, this capability introduces an important limitation: the model's performance is contingent upon the presence of the specific fiducial marker used during training. Experimental results indicate that variations in marker size or pattern would likely render the model ineffective, as it has not learned to generalize beyond the predefined marker characteristics. This dependency represents a significant constraint for practical applications requiring flexible marker configurations.

Table 5. Average MAPE on different distances to the camera and the marker

Distance (cm)	Without Fiducial marker (%)	With Fiducial marker (%)
10	12.76	9.21
15	10.61	8.38
20	8.10	7.51
25	15.68	7.81
30	30.46	9.73

4. Practical deployment challenges

While the proposed framework demonstrates promising accuracy in controlled laboratory settings, its deployment in real-world environments introduces several technical and ethical challenges. Addressing these challenges is essential for ensuring both practical feasibility and responsible use, particularly in safety-critical domains such as predictive maintenance and structural health monitoring.

4.1. Maker dependence

A key limitation is the model's reliance on fiducial markers for accurate motion estimation. As demonstrated in Table 5, prediction accuracy declined significantly when the marker was occluded, particularly at increased camera-to-object distances. This reliance constrains the system's applicability in real-world scenarios where marker placement may be impractical or inconsistent. A potential solution involves employing feature detection on natural structures to serve as references for calibrating the distance-per-pixel parameter [37]. However, this approach may introduce challenges, such as elevated computational costs, rendering it impractical for real-time applications. Alternative approaches, such as stereo camera setups or single-camera systems with distance measurement sensors, offer potential solutions [38]. Given computational constraints, future efforts will first prioritize natural feature-based methods, which offer a balance between generalization and real-time feasibility, before exploring hardware-intensive solutions.

4.2. Data biases

The dataset used in this study was generated from a single-object, laboratory-scale setup with synthetic augmentation to increase frequency diversity. While effective for proof-of-concept validation, this introduces biases toward specific vibration patterns, lighting levels, and object characteristics. Such biases may reduce model robustness when applied to multi-component machinery or noisy industrial environments, where vibration signals are more complex and varied.

To address this, future work will expand the dataset in two stages: (1) collecting data from machines of varied shapes, materials, and operating conditions in controlled laboratory settings, and (2) collaborating with Sabah Electricity Utility Company to acquire data from real industrial systems. Additionally, optimizing hyperparameters, such as learning rates, length of sequence, and model architectures, will enhance performance and adaptability to diverse datasets [39-41]. To mitigate potential long-term data limitations, integrating recurrent neural networks (RNNs) into the model will enable stream learning, allowing continuous adaptation to new vibration patterns and environmental conditions [42]. Rigorous evaluation on diverse, unseen datasets from real-world applications, using metrics like mean absolute error for displacement and velocity, will be essential to quantify performance gaps and ensure robust generalization across complex scenarios.

4.3. Real-time application

In terms of performance, the model achieved an average inference time of 9.95 ms on workstation hardware but has yet to be benchmarked on embedded devices. Future work will focus on deployment to the Raspberry Pi 5, beginning with TensorFlow Lite (TFLite) conversion and quantization to reduce model size and latency. Lightweight architectures such as MobileNetV3 or EfficientNet-Lite for spatial feature extraction, combined with temporal modules like Temporal Convolutional Networks (TCNs), may further improve efficiency [43].

The goal is to approach real-time inference on edge hardware (e.g., <10 ms per frame), enabling continuous monitoring under typical industrial workloads. Profiling tools such as TensorFlow Lite Benchmark, perf, and hardware energy sensors will be used to quantify trade-offs in latency, power consumption, and accuracy. Multi-objective optimization methods (e.g., NSGA-II) may also be applied to balance competing requirements for accuracy and efficiency [44].

4.4. Frequency predictions

Compared to other vibration metrics, frequency predictions exhibited higher error, as shown in Tables 2 and 3. This is partly due to limited dataset variability: only four dominant natural frequencies were present in the collected data, restricting the model's exposure to diverse vibration modes.

To improve temporal modeling, hybrid architectures such as CNN–LSTM, GRUs, or transformer-based models will be explored. CNN–LSTM and GRU architectures are well-suited for capturing sequential dependencies while remaining computationally efficient, making them candidates for edge deployment [45-47]. Transformer-based approaches may further enhance robustness by modeling long-range dependencies across frame sequences [48]. Future evaluations will benchmark these architectures against the current (2+1)D CNN backbone, focusing on accuracy, generalization, and real-time performance. Future work will explore these models, comparing their performance against the baseline model in terms of accuracy and computational efficiency, as outlined in the preceding subsection.

4.5. Safety-critical considerations

Beyond technical limitations, broader implications must also be considered, particularly when deploying ML-driven monitoring in safety-critical domains. The reliability of predictions may degrade under uncontrolled conditions such as occlusions, extreme lighting, or motion blur, potentially exceeding reported error margins. In high-stakes applications such as monitoring bridges, aircraft, or power grid infrastructure, such degradation could compromise timely fault detection.

Misclassifications pose additional risks. False negatives may allow undetected faults to progress into catastrophic failures, while false positives may trigger unnecessary maintenance, leading to downtime and economic losses. These risks highlight the need for human-in-the-loop oversight, with ML predictions serving as advisory tools supported by redundancy, fail-safe mechanisms, and cross-validation against conventional sensors.

Furthermore, ethical considerations arise from the use of video-based monitoring. Although this framework is intended for machine vibration analysis, deployment in shared environments could inadvertently capture sensitive or personally identifiable information. Robust privacy safeguards, including anonymization protocols and compliance with data protection regulations, are therefore essential.

To mitigate these risks, future development should prioritize dataset diversification to reduce bias, integrate uncertainty quantification to support decision-making, and adopt ethical deployment frameworks that emphasize transparency, accountability, and privacy. These measures will ensure that the proposed framework evolves into a responsible and trustworthy solution for predictive maintenance in safety-critical systems.

5. Conclusions

This study presented and validated a vision-based deep learning framework for non-contact vibration measurement using a Four-Output Regression (2+1)D CNN architecture combined with optical flow preprocessing. The model demonstrated the ability to predict acceleration, velocity, displacement, and frequency from video input with high accuracy, achieving an average MAPE of 7.51 %. Computational efficiency was improved by Region of Interest (ROI) cropping and multi-section frame extraction, reducing the required number of input frames from 30 to 15. Dataset augmentation further improved performance under varied lighting conditions and frequency ranges, reducing MAPE for frequency prediction by 35 %.

Despite these encouraging results, several deployment challenges persist. The model's reliance on fiducial markers restricts its practical applicability, and its generalization to more complex systems is limited by the current dataset, which is confined to single-object scenarios. Moreover, the performance on embedded platforms remains untested, and robustness in uncontrolled environments has yet to be fully established.

Future work will focus on expanding the dataset using industrial-scale equipment, integrating multi-sensor data for improved robustness, exploring marker-less tracking approaches, and optimizing the model for edge-device deployment. These advancements are expected to support the development of a scalable, non-invasive, and cost-effective solution for predictive maintenance and structural health monitoring, offering a viable alternative to traditional contact-based sensing systems.

Acknowledgements

This study was supported by the Malaysian Ministry of Higher Education (MOHE) Fundamental Research Grant Scheme (FRGS/1/2023/TK10/UMS/02/1).

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Author contributions

Harold Harrison: conceptualization, data curation, original draft preparation. Mazlina Mamat: funding acquisition, methodology, review and editing. Farrah Wong: review, resources. Hoe Tung Yew: review, visualization. Racheal Lim: data validation. Wan Mimi Diyana Wan Zaki: funding acquisition, resources.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] L. Lin, D. Wang, S. Zhao, L. Chen, and N. Huang, "Power quality disturbance feature selection and pattern recognition based on image enhancement techniques," *IEEE Access*, Vol. 7, pp. 67889–67904, Jan. 2019, https://doi.org/10.1109/access.2019.2917886

[2] F. Mevissen and M. Meo, "A review of NDT/Structural health monitoring techniques for hot gas components in gas turbines," *Sensors*, Vol. 19, No. 3, p. 711, Feb. 2019, https://doi.org/10.3390/s19030711

[3] R. Soman, K. Balasubramaniam, A. Golestani, M. Karpiński, and P. Malinowski, "A two-step guided waves based damage localization technique using optical fiber sensors," *Sensors*, Vol. 20, No. 20, p. 5804, Oct. 2020, https://doi.org/10.3390/s20205804

[4] A. Taghizadeh-Alisaraei and A. Mahdavian, "Fault detection of injectors in diesel engines using vibration time-frequency analysis," *Applied Acoustics*, Vol. 143, pp. 48–58, Jan. 2019, https://doi.org/10.1016/j.apacoust.2018.09.002

[5] H. Zhao et al., "A highly sensitive triboelectric vibration sensor for machinery condition monitoring," *Advanced Energy Materials*, Vol. 12, No. 37, Aug. 2022, https://doi.org/10.1002/aenm.202201132

[6] M. Mihaila, I. Blanari, V. Goanta, and P. D. Barsanescu, "Finite element analysis of a modified weigh-in-motion sensor," *IOP Conference Series: Materials Science and Engineering*, Vol. 1262, No. 1, p. 012052, Oct. 2022, https://doi.org/10.1088/1757-899x/1262/1/012052

[7] C. Deng and L. Hong, "Research on a non-contact magnetically coupled vibration sensor of fiber grating," in *3rd International Conference on Optoelectronic Information and Functional Materials (OIFM 2024)*, p. 88, Jul. 2024, https://doi.org/10.1117/12.3034332

[8] R. Wang et al., "Non-contact dynamic capacity-increasing of overhead conductor based on cooling tester (CT)," in *E3S Web of Conferences*, Vol. 185, p. 01078, Sep. 2020, https://doi.org/10.1051/e3sconf/202018501078

[9] P. Garg et al., "Measuring transverse displacements using unmanned aerial systems laser doppler vibrometer (UAS-LDV): development and field validation," *Sensors*, Vol. 20, No. 21, p. 6051, Oct. 2020, https://doi.org/10.3390/s20216051

[10] M. Muramatsu, S. Uchida, and Y. Takahashi, “Noncontact detection of concrete flaws by neural network classification of laser doppler vibrometer signals,” *Engineering Research Express*, Vol. 2, No. 2, p. 025017, May 2020, https://doi.org/10.1088/2631-8695/ab8ba4

[11] D. Maillard, A. de Pastina, A. M. Abazari, and L. G. Villanueva, “Avoiding transduction-induced heating in suspended microchannel resonators using piezoelectricity,” *Microsystems and Nanoengineering*, Vol. 7, No. 1, Apr. 2021, https://doi.org/10.1038/s41378-021-00254-1

[12] V. I. Moreno-Oliva et al., “Vibration measurement using laser triangulation for applications in wind turbine blades,” *Symmetry*, Vol. 13, No. 6, p. 1017, Jun. 2021, https://doi.org/10.3390/sym13061017

[13] J. Goszczak, G. Mitukiewicz, and D. Batory, “The influence of material’s surface modification on the structure’s dynamics-initial test results,” in *Journal of Physics: Conference Series*, Vol. 2698, No. 1, p. 012010, Feb. 2024, https://doi.org/10.1088/1742-6596/2698/1/012010

[14] M. Romanssini, P. C. C. de Aguirre, L. Compassi-Severo, and A. G. Girardi, “A review on vibration monitoring techniques for predictive maintenance of rotating machinery,” *Eng*, Vol. 4, No. 3, pp. 1797–1817, Jun. 2023, https://doi.org/10.3390/eng4030102

[15] R. Wu, Y. Li, and S. Zhang, “Strain fields measurement using frequency domain Savitzky-Golay filters in digital image correlation,” *Measurement Science and Technology*, Vol. 34, No. 9, p. 095115, Sep. 2023, https://doi.org/10.1088/1361-6501/acda53

[16] A. Molina-Viedma, E. López-Alba, L. Felipe-Sesé, and F. Díaz, “Full-field operational modal analysis of an aircraft composite panel from the dynamic response in multi-impact test,” *Sensors*, Vol. 21, No. 5, p. 1602, Feb. 2021, https://doi.org/10.3390/s21051602

[17] Y. Wei et al., “Interferometric-scale full-field vibration measurement by a combination of digital image correlation and laser vibrometer,” *Optics Express*, Vol. 32, No. 12, p. 20742, Jun. 2024, https://doi.org/10.1364/oe.521211

[18] S. Yan and Z. Zhang, “3D mode shapes characterization under hammer impact using 3D-DIC and phase-based motion magnification,” *Engineering Research Express*, Vol. 6, No. 3, p. 035544, Sep. 2024, https://doi.org/10.1088/2631-8695/ad6e53

[19] S. Cao, J. Yan, H. Nian, and C. Xu, “Full-field out-of-plane vibration displacement acquisition based on speckle-projection digital image correlation and its application in damage localization,” *International Journal of Mechanical System Dynamics*, Vol. 2, No. 4, pp. 363–373, Nov. 2022, https://doi.org/10.1002/msd2.12055

[20] M. Uusinoka, J. Haapala, and A. Polojärvi, “Deep learning-based optical flow in fine-scale deformation mapping of sea ice dynamics,” *Geophysical Research Letters*, Vol. 52, No. 2, Jan. 2025, https://doi.org/10.1029/2024gl112000

[21] L. Su, H. Huang, L. Qin, and W. Zhao, “Transformer vibration detection based on YOLOv4 and optical flow in background of high proportion of renewable energy access,” *Frontiers in Energy Research*, Vol. 10, Feb. 2022, https://doi.org/10.3389/fenrg.2022.764903

[22] Z. Peng, M. Liu, Z. Wang, W. Liu, and X. Wang, “Phase-based optical flow method with optimized parameter settings for enhancing displacement measurement adaptability,” *Open Journal of Applied Sciences*, Vol. 14, No. 5, pp. 1165–1184, Jan. 2024, https://doi.org/10.4236/ojapps.2024.145075

[23] Z. Peng, X. Wang, W. Liu, Z. Wang, and M. Liu, “Assessment of displacement measurement performance using phase-based optical flow with the HalfOctave pyramid,” in *Eleventh International Symposium on Precision Mechanical Measurements*, p. 49, Sep. 2024, https://doi.org/10.1117/12.3032543

[24] T. Manabe and Y. Shibata, “Real-time image-based vibration extraction with memory-efficient optical flow and block-based adaptive filter,” *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, Vol. E106.A, No. 3, pp. 504–513, Mar. 2023, https://doi.org/10.1587/transfun.2022vlp0009

[25] M. Civera, L. Zanotti Fragonara, and C. Surace, “An experimental study of the feasibility of phase-based video magnification for damage detection and localisation in operational deflection shapes,” *Strain*, Vol. 56, No. 1, Jan. 2020, https://doi.org/10.1111/str.12336

[26] Y. Bai, H. Sezen, A. Yilmaz, and R. Qin, “Bridge vibration measurements using different camera placements and techniques of computer vision and deep learning,” *Advances in Bridge Engineering*, Vol. 4, No. 1, Nov. 2023, https://doi.org/10.1186/s43251-023-00105-1

[27] J. Nam and J. Kang, “Classification of chaotic squeak and rattle vibrations by CNN using recurrence pattern,” *Sensors*, Vol. 21, No. 23, p. 8054, Dec. 2021, https://doi.org/10.3390/s21238054

[28] C. Du et al., “Unmanned aerial vehicle rotor fault diagnosis based on interval sampling reconstruction of vibration signals and a one-dimensional convolutional neural network deep learning method,” *Measurement Science and Technology*, Vol. 33, No. 6, p. 065003, Jun. 2022, https://doi.org/10.1088/1361-6501/ac491e

[29] J. Siłka, M. Wieczorek, and M. Woźniak, “Recurrent neural network model for high-speed train vibration prediction from time series,” *Neural Computing and Applications*, Vol. 34, No. 16, pp. 13305–13318, Jan. 2022, https://doi.org/10.1007/s00521-022-06949-4

[30] S. Jain, G. Seth, A. Paruthi, U. Soni, and G. Kumar, “Synthetic data augmentation for surface defect detection and classification using deep learning,” *Journal of Intelligent Manufacturing*, Vol. 33, No. 4, pp. 1007–1020, Nov. 2020, https://doi.org/10.1007/s10845-020-01710-x

[31] P. Nardelli, J. C. Ross, and R. San José Estépar, “Generative-based airway and vessel morphology quantification on chest CT images,” *Medical Image Analysis*, Vol. 63, p. 101691, Jul. 2020, https://doi.org/10.1016/j.media.2020.101691

[32] R. Yang et al., “CNN-LSTM deep learning architecture for computer vision-based modal frequency detection,” *Mechanical Systems and Signal Processing*, Vol. 144, p. 106885, Oct. 2020, https://doi.org/10.1016/j.ymssp.2020.106885

[33] B. Sbaiti, J. D. Schultz, K. A. Parker, and D. N. Beratan, “Machine learning for video classification enables quantifying intermolecular couplings from simulated time-evolved multidimensional spectra,” *The Journal of Physical Chemistry Letters*, Vol. 16, No. 19, pp. 4707–4714, May 2025, https://doi.org/10.1021/acs.jpclett.5c00588

[34] T.-C. Chan, C.-C. Chang, A. Ullah, and H.-H. Lin, “Study on kinematic structure performance and machining characteristics of 3-axis machining center,” *Applied Sciences*, Vol. 13, No. 8, p. 4742, Apr. 2023, https://doi.org/10.3390/app13084742

[35] B. Rafael, A.A. Muhammad Zacky, and K. Irwan, “Electricity consumption prediction in oil and gas equipment service and maintenance workshops using RNN LSTM,” in *E3S Web of Conferences*, Vol. 426, p. 02089, Sep. 2023, https://doi.org/10.1051/e3sconf/202342602089

[36] D. Hirahara, E. Takaya, M. Kadowaki, Y. Kobayashi, and T. Ueda, “Effect of the pixel interpolation method for downsampling medical images on deep learning accuracy,” *Journal of Computer and Communications*, Vol. 9, No. 11, pp. 150–156, Jan. 2021, https://doi.org/10.4236/jcc.2021.911010

[37] M. A. Kuddus, J. Li, H. Hao, C. Li, and K. Bi, “Target-free vision-based technique for vibration measurements of structures subjected to out-of-plane movements,” *Engineering Structures*, Vol. 190, pp. 210–222, Jul. 2019, https://doi.org/10.1016/j.engstruct.2019.04.019

[38] J. Li et al., “LiDAR-assisted UAV stereo vision detection in railway freight transport measurement,” *Drones*, Vol. 6, No. 11, p. 367, Nov. 2022, https://doi.org/10.3390/drones6110367

[39] E.-S. M. El-Kenawy et al., “Feature selection in wind speed forecasting systems based on meta-heuristic optimization,” *PLOS ONE*, Vol. 18, No. 2, p. e0278491, Feb. 2023, https://doi.org/10.1371/journal.pone.0278491

[40] A. A. Alhussan et al., “A binary waterwheel plant optimization algorithm for feature selection,” *IEEE Access*, Vol. 11, pp. 94227–94251, Jan. 2023, https://doi.org/10.1109/access.2023.3312022

[41] E.-S. M. El-Kenawy, A. Ibrahim, S. Mirjalili, Y.-D. Zhang, S. Elnazer, and R. M. Zaki, “Optimized ensemble algorithm for predicting metamaterial antenna parameters,” *Computers, Materials and Continua*, Vol. 71, No. 3, pp. 4989–5003, Jan. 2022, https://doi.org/10.32604/cmc.2022.023884

[42] F. K. Karim, D. S. Khafaga, M. M. Eid, S. K. Towfek, and H. K. Alkahtani, “A novel bio-inspired optimization algorithm design for wind power engineering applications time-series forecasting,” *Biomimetics*, Vol. 8, No. 3, p. 321, Jul. 2023, https://doi.org/10.3390/biomimetics8030321

[43] D. K.-H. Lai et al., “Transformer models and convolutional networks with different activation functions for swallow classification using depth video data,” *Mathematics*, Vol. 11, No. 14, p. 3081, Jul. 2023, https://doi.org/10.3390/math11143081

[44] N. Khodadadi, L. Abualigah, E.-S. M. El-Kenawy, V. Snasel, and S. Mirjalili, “An archive-based multi-objective arithmetic optimization algorithm for solving industrial engineering problems,” *IEEE Access*, Vol. 10, pp. 106673–106698, Jan. 2022, https://doi.org/10.1109/access.2022.3212081

[45] M. Cao, R. Yao, J. Xia, K. Jia, and H. Wang, “LSTM attention neural-network-based signal detection for hybrid modulated faster-than-nyquist optical wireless communications,” *Sensors*, Vol. 22, No. 22, p. 8992, Nov. 2022, https://doi.org/10.3390/s22228992

[46] L. Wu, C. Kong, X. Hao, and W. Chen, “A short-term load forecasting method based on GRU-CNN hybrid neural network model,” *Mathematical Problems in Engineering*, Vol. 2020, pp. 1–10, Mar. 2020, https://doi.org/10.1155/2020/1428104

[47] Y. Xie, J. Zhao, B. Qiang, L. Mi, C. Tang, and L. Li, "Attention mechanism-based CNN-LSTM model for wind turbine fault prediction using SSN ontology annotation," *Wireless Communications and Mobile Computing*, Vol. 2021, No. 1, p. 66275, Mar. 2021, https://doi.org/10.1155/2021/6627588

[48] B. Zhang, L. Liu, M. H. Phan, Z. Tian, C. Shen, and Y. Liu, "SegViT v2: exploring efficient and continual semantic segmentation with plain vision transformers," *International Journal of Computer Vision*, Vol. 132, No. 4, pp. 1126–1147, Oct. 2023, https://doi.org/10.1007/s11263-023-01894-8

Harold Harrison achieved his B.Eng. degree in electronic engineering (computer) from University Malaysia Sabah in 2023. At present, he serves as a research assistant affiliated with the Applied Predictive Analytics Research Group within the Electronic Engineering (Computer) Program at the Faculty of Engineering, University of Malaysia Sabah. His academic pursuits are focused on the dynamic fields of image processing, pattern recognition, and the Internet of Things (IoT). His engagement in these areas reflects a keen interest in cutting-edge technologies and their applications.

Mazlina Mamat received her B.Eng. and M.Sc. from University Sains Malaysia (2000, 2003) and Ph.D. in microengineering and nanoelectronics from The National University of Malaysia (2013). She is a Senior Lecturer at University of Malaysia Sabah and a registered Professional Engineer with the Board of Engineers Malaysia. She has co-authored over 40 publications, including books and journals. Her research interests include sensors, signal processing, machine learning, and AI.

Farrah Wong received the B.Eng, Master and Doctor of Philosophy degrees from University of Malaysia Sabah in 1999, 2001 and 2004 respectively. She has joined the Faculty of Engineering, University of Malaysia Sabah as a lecturer since 2004. Farrah Wong is a registered professional engineer with the board of engineers Malaysia. Her research interests are in intelligent and vision systems.

Hoe Tung Yew received his B.Eng. from the University of Lincoln (2003), M.Sc. from Northumbria University (2004), and Ph.D. in biomedical engineering from University of Technology Malaysia. He is a Senior Lecturer at University of Malaysia Sabah, a Chartered Engineer (IET, U.K.), and a Professional Engineer (Malaysia). He has published over 50 peer-reviewed papers. His research focuses on wireless communications, including vertical handover, sensor networks, and IoT.

Racheal Lim holds a B.Eng. in electrical and electronic engineering from University of Malaysia Sabah (2001). She is the chief transmission officer at Sabah Electricity Sdn Bhd with 21 years of experience in power utilities. She is a certified 275kV electrical service and competent engineer by the energy commission, and a registered professional engineer and technologist. A corporate member of IEM, her interests include grid reliability, digitalization, and predictive maintenance.

Wan Mimi Diyana is an Associate Professor specializing in biomedical and bioinformatics engineering, focusing on smart systems, image processing, and IoT-driven healthcare. Her research emphasizes AI-based medical imaging for ocular diseases such as cataract and diabetic retinopathy. She holds a Ph.D. in engineering from Multimedia University and has published over 150 papers. She actively contributes to professional organizations and has secured major national and international research grants in her field.