4th International Civil Engineering & Architecture Conference
17-19 May 2025, Trabzon, Türkiye

Advanced deep learning models for anomaly detection in masonry structures

Kemal Hacıefendioğlu*[1], Ali Motamedi[1], Tunahan Aslan[1], Cemile Duman[2], Ayşecan Bostan[1]

[1]*Karadeniz Technical University, Department of Civil Engineering, 61080 Trabzon, Türkiye*
[2] *Gazi University, Faculty of Engineering, Department of Civil Engineering, Ankara, Türkiye*

Abstract. The detection of structural anomalies is essential for safeguarding the integrity and longevity of buildings with cultural and historical significance, such as masonry minarets. This study explores the use of Convolutional Neural Networks (CNNs) for forecasting vibration responses and identifying patterns indicative of potential structural damage. Vibration data collected from a minaret under both undamaged and controlled damage scenarios are used to train and evaluate a CNN model designed to extract spatial features from time-series signals. The model demonstrates strong performance in distinguishing normal structural behavior from anomalous patterns, offering high accuracy and sensitivity in damage detection. In addition to its predictive capabilities, the CNN approach proves to be computationally efficient and suitable for real-time monitoring applications. The findings highlight the potential of deep learning–based methods, particularly CNNs, to enhance non-contact structural health monitoring strategies, providing a reliable and scalable solution for the preservation of historic structures.

Keywords: Historic masonry structures; Optical flow; Convolutional Neural Networks (CNN); Anomaly detection; Non-contact measurement

1. Introduction

Historic masonry structures, such as minarets, are significant cultural heritage assets that often experience deterioration over time due to aging, environmental conditions, and dynamic forces such as earthquakes (Sabia, Demarie et al., 2022). Maintaining their structural integrity is essential due to their historical and cultural importance. Traditional structural health monitoring (SHM) methods are often invasive and limited by practical challenges and the risk of further damage (Dong and Catbas, 2021). As a result, non-contact and non-intrusive methods, especially video-based SHM, have become increasingly popular for being safe, cost-effective, and efficient (Shariati, Schumacher et al., 2014).

Video-based SHM uses advanced computer vision, image processing, and machine learning (ML) techniques to detect structural damage by analyzing video data. This method provides significant benefits in scalability, accuracy, and safety, allowing continuous monitoring without physical contact (Ghandil, Dabanli et al., 2021; Xiong, 2021). Although video-based SHM has been widely studied, research focused specifically on historic masonry minarets are limited. These structures present unique challenges for monitoring due to their complex geometry, varied materials, and cultural significance.

Recent progress in video-based SHM has highlighted its effectiveness in monitoring structural dynamics and detecting gradual damage in historic buildings. Digital video imaging, in particular, has proven suitable for non-contact dynamic assessment, offering a practical solution for historic masonry structures where traditional sensors are often impractical or pose a risk to the structure (Ghandil, Dabanli et al., 2021). Graph-based video processing techniques have emerged as effective tools for early warning systems in historic masonry structures. By analyzing key deformation patterns, these methods can help predict potential structural failures, offering a means to protect valuable heritage sites from severe damage (Fioriti, Cataldo et al., 2024).

Automated damage detection in masonry using video imaging shortens inspection times and improves safety by removing the need for inspectors to enter hazardous areas (Sangirardi, Altomare et al., 2022). Moreover, advancements in machine learning have greatly enhanced the effectiveness of video-based SHM. Sophisticated algorithms that detect subtle structural changes in video data support long-term monitoring and predictive

*Corresponding author, E-mail: kemalhaciefendioglu@ktu.edu.tr

https://doi.org/10.31462/icearc2025_ce_eqe_529

maintenance, which are especially important for historic masonry buildings exposed to seismic activity and other dynamic forces (Aktürk, Aras et al., 2022; Farrar, Dervilis et al., 2025). Although video-based methods have advanced significantly, their use in monitoring historic masonry minarets has received limited attention. This study aims to bridge that gap by presenting a tailored video-based monitoring framework that considers the unique structural traits and preservation demands of minarets.

This study presents a novel integration of advanced optical flow techniques with deep learning (DL) architectures, specifically CNN model, to capture the distinctive structural behavior of masonry minarets and detect anomalies. The key contributions are: (a) directly addressing the monitoring challenges arising from the complex geometry and material variability of historic masonry minarets, (b) introducing a hybrid Autoencoder-PCA method to efficiently reduce dimensionality and improve anomaly detection by isolating dominant patterns in displacement time-series data, and (c) conducting a detailed analysis of CNN model to evaluate its performance and suitability for video-based monitoring of heritage structures.

2. Theoretical background

2.1. Optical flow methods

Optical flow refers to the apparent motion of objects or surfaces in a scene due to movement between the camera and the environment. In structural health monitoring (SHM), it enables real-time, non-contact tracking of displacements, deformations, and vibrations.

The Lucas-Kanade method is well-suited for tracking small displacements, whereas its extension through the pyramid approach improves reliability in cases involving larger movements (Lucas and Kanade, 1981; Bouguet, 2001; Baker and Matthews, 2004).

2.2. Hybrid autoencoder-PCA

Autoencoders (AEs) are neural networks that use unsupervised learning to extract efficient, lower-dimensional representations of data. They consist of two main parts: an encoder that compresses the input and a decoder that reconstructs it, enabling meaningful features learning from the original data.

2.3. Convolutional neural networks (CNNs)

Convolutional Neural Networks (CNNs) play a pivotal role in deep learning, especially for analyzing visual data such as images and videos. Unlike conventional neural networks that rely on manually selected features, CNNs automatically learn spatial hierarchies of features from raw input, enabling efficient processing of high-dimensional and complex datasets. Their architecture, inspired by the structure of the visual cortex, consists of multiple layers that progressively extract informative patterns relevant to tasks like classification and detection (Fig. 1). The process begins with a convolutional layer that applies learnable filters (kernels) to the input data, extracting features that range from basic edges and textures to more abstract patterns in deeper layers. An activation function—typically the Rectified Linear Unit (ReLU)—is then used to introduce nonlinearity, allowing the network to model complex relationships. Pooling layers follow this step, reducing the spatial dimensions of the feature maps while retaining essential information, which helps lower computational cost and control overfitting. After several convolution and pooling operations, fully connected layers flatten and integrate the extracted features for classification or regression. Each neuron in these layers connects to all neurons in the next layer, enabling high-level reasoning. A SoftMax layer is typically used at the output to convert raw scores into class probabilities, allowing the model to produce interpretable predictions in visual recognition tasks (Simonyan and Zisserman, 2014). CNNs are particularly well-suited for image processing and pattern recognition tasks. Their primary strength lies in their ability to automatically extract hierarchical features from raw data, eliminating the need for manual feature engineering. This makes CNNs highly effective in domains that involve complex visual data.

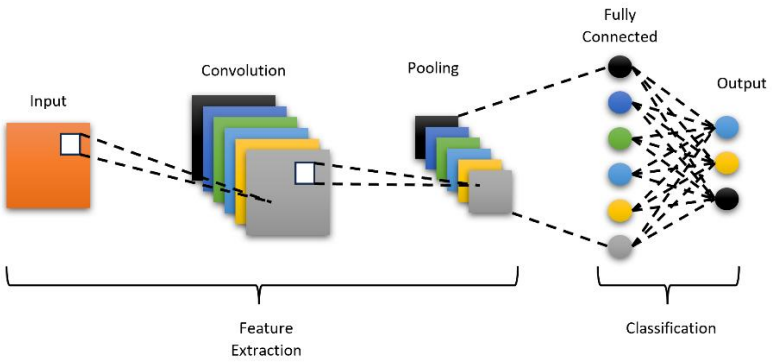


Fig. 1. The structure of a CNN

CNNs have revolutionized many fields, especially those that require the analysis of complex visual data. Their ability to automatically learn features from raw data has made them indispensable in tasks such as image classification, object detection, and medical image analysis. As technology continues to evolve, the applications of CNNs will likely expand even further, playing a critical role in innovations across a wide range of industries. Their effectiveness in understanding complex patterns, coupled with their efficiency, ensures that CNNs will remain a foundational tool in the field of deep learning (Szegedy, Liu et al. 2015).

3. Anomaly assessment methodology

The proposed approach combines advanced image processing with unsupervised learning to evaluate the structural condition of a portable minaret model. Four cameras, positioned at key locations, record high-resolution video of markers placed on the minaret's surface, allowing accurate measurement of both horizontal and vertical displacements under intact and damaged states. Optical flow techniques are used to calculate these displacements, which are then translated from pixel units into real-world values through calibration using known marker dimensions. The resulting displacement time series is used for feature extraction, where a hybrid Autoencoder-PCA model captures the latent patterns of healthy behavior. In parallel CNN architecture leverage both spatial and temporal features to improve anomaly detection. A reconstruction error-based damage index, computed in an unsupervised manner, identifies anomalies by comparing original and reconstructed signals, with detection refined through thresholding and segment-level analysis. This integrated framework enables continuous monitoring and clear visualization of structural damage progression over time.

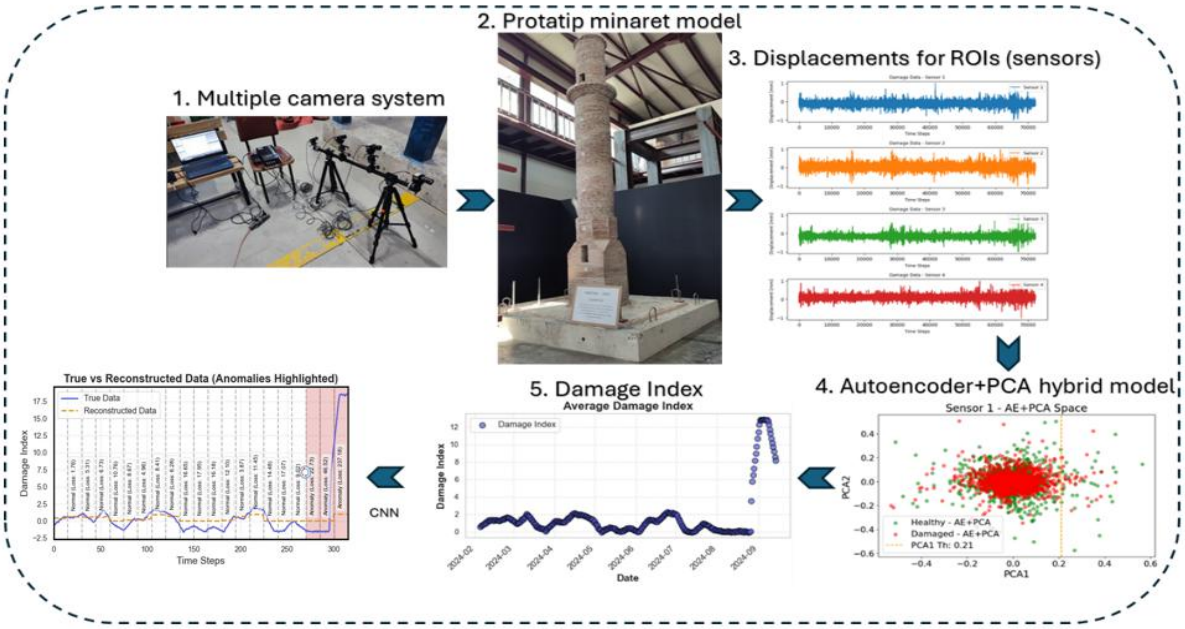


Fig. 2. Steps of proposed damage strategy

Figure 2 illustrates the workflow of the proposed methodology, with each step outlined as follows:

Step 1: Multi-Camera System and Prototype Minaret Model: A portable minaret model was equipped with four cameras positioned strategically to focus on specific markers placed on the surface of the structure. This setup enabled each camera to independently record video data, resulting in four separate and complementary data streams.

Step 2: Displacements for ROI. Each video stream is processed using the optical flow method to compute pixel movements between consecutive frames, capturing horizontal displacements of the minaret. Marker positions and their motion patterns are analysed to distinguish displacements at different locations on the structure. The resulting pixel-based motion vectors are extracted and stored, creating a raw dataset for further analysis.

The pixel-based motion vectors derived from optical flow are then converted into real-world units, such as millimeters or centimeters. This conversion relies on scaling factors determined through the known dimensions of the markers used in the experimental setup, combined with camera calibration parameters to ensure accurate spatial measurements.

These parameters allow for precise conversion of pixel displacements into actual physical distances. The resulting horizontal and vertical displacement values are recorded as time series data for multiple points on the minaret, creating a detailed dataset for subsequent analysis.

Step 3: Feature extraction and model training. To facilitate damage detection, the displacement time series data are transformed into a more compact and informative representation using a hybrid Autoencoder-PCA model. The autoencoder compresses and reconstructs the input signals, capturing their key characteristics. PCA is then applied to further reduce dimensionality and highlight dominant patterns. Importantly, only data from the undamaged (healthy) state are used during training, enabling the model to learn a latent representation that characterizes normal structural behaviour.

Step 4: Unsupervised reconstruction error-based structural damage index. The model is initially trained on the undamaged (reference) dataset of the minaret. Once trained, it is applied to new measurements to calculate a damage index based on reconstruction error. By comparing how well the model reconstructs current data relative to the reference state, deviations indicating potential damage are identified. Tracking these indices over time allows for visualization of damage progression throughout the monitoring period.

DL models are designed to compress input data and reconstruct it with minimal loss. The reconstruction error quantifies how effectively the model can reproduce the input.

4. Experimental implementation

4.1. Controlled damage scenarios

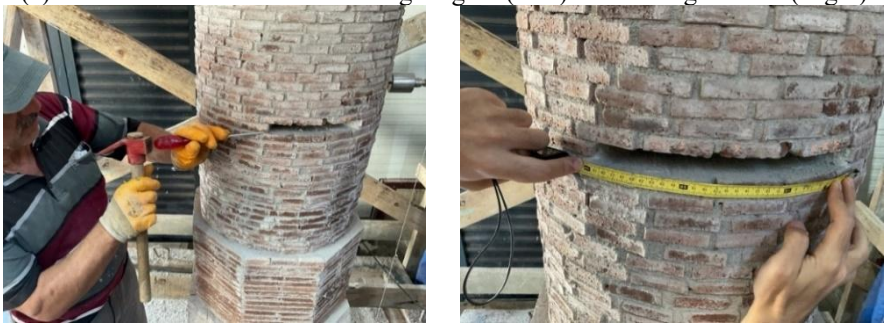
The damage was introduced by selectively removing bricks, a method chosen to observe changes in the structure's dynamic behavior. The targeted damage area was located on the brick wall section aligned with the lowest-level accelerometer, where seven bricks were removed to enable effective 2D displacement measurement. This damage zone was situated roughly 35 cm above the transition section, at a height of 145 cm from the base of the minaret. The procedure used for brick removal is illustrated in Figure 3 and detailed in the following paragraphs:

- Step 1 (Damage-1). The initial damage phase involved the removal of two out of the seven designated bricks, as shown in Figure 3a.
- Step 2 (Damage-2). In the next phase, four additional bricks located next to the initially removed ones were taken out, with two on each side. This resulted in a damaged region measuring approximately 30 cm in width and 3 cm in height, as shown in Figure 3b.
- Step 3 (Damage-3). As in the previous steps, one additional brick was removed from the upper brick layer, bringing the total number of removed bricks to five. Displacement measurements using the camera system were performed after each removal, as shown in Figure 3c.
- Step 4 (Damage-4). In the final stage, three additional bricks were removed from the upper brick layer, resulting in a total of seven bricks being extracted. After each removal, camera-based displacement measurements were repeated, as illustrated in Figure 3d.

Throughout the damage induction process, camera-based measurements were consistently recorded to monitor and evaluate the structural response of the minaret model.



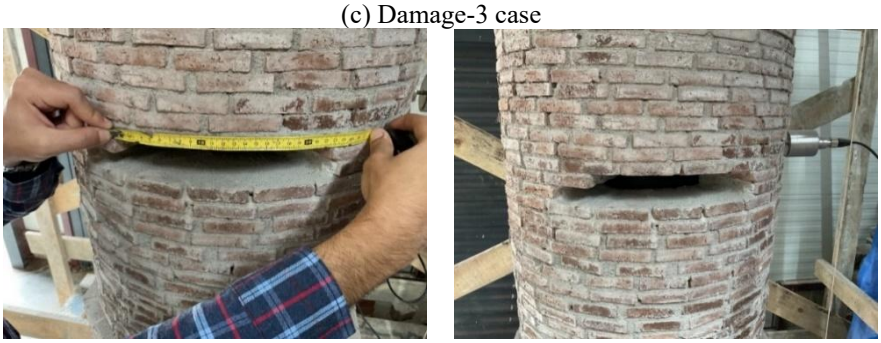
(a) Location of the controlled damage region (Left) and Damage-1 case (Right)



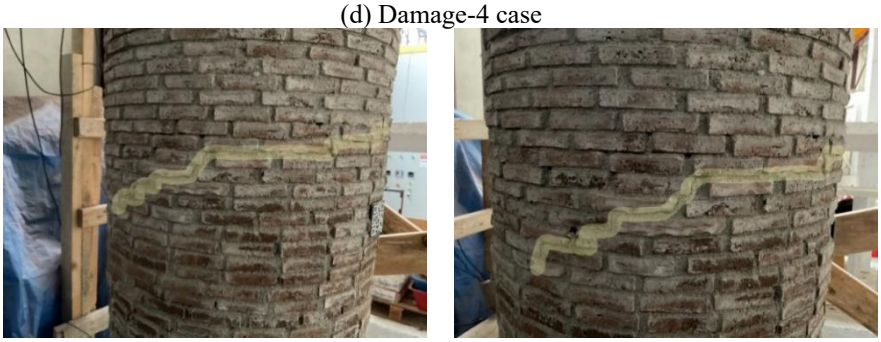
(b) Damage-2 case



(c) Damage-3 case



(d) Damage-4 case



(e) Cracks formed after damage

Fig. 3. Controlled damage scenarios considered in the experimental study

4.2. Data collection and analysis

The monitoring was carried out using predefined regions of interest (ROIs) on the scaled minaret model. A template-matching-based sensing system continuously tracked the motion of selected points to collect displacement data. Measurements were performed using a multi-camera setup, and the data from each camera were analyzed comparatively. Figure 4 illustrates the placement of the template markers and the camera configuration. In this setup, the most distant camera was aimed at the top section of the model, specifically the honeycomb structure, while the other cameras were directed toward the podium, transition zone, and shaft at predefined angles. This arrangement ensured complete high-resolution coverage of all critical regions. The structured positioning allowed each camera to capture motion at different levels, ensuring synchronized and consistent data collection, and enabling accurate analysis of structural behavior and dynamic response.



Fig. 4. Template markers on the minaret and camera layout

During the monitoring process, periodic displacement measurements were analyzed to identify variations in the structure's dynamic behavior. The reference dataset, representing the undamaged state, includes 216,433 data points collected from four distinct measurement locations, as shown in Figure 5. This figure also displays the time-dependent displacement trends for both the training and test sets under undamaged conditions. After introducing controlled damage to the minaret model, measurements were repeated at the same four locations. For instance, the displacement data corresponding to the Damage-4 condition are detailed in Figure 6. These evaluations offer valuable insights into the model's dynamic response following damage and support the assessment of its structural stability.

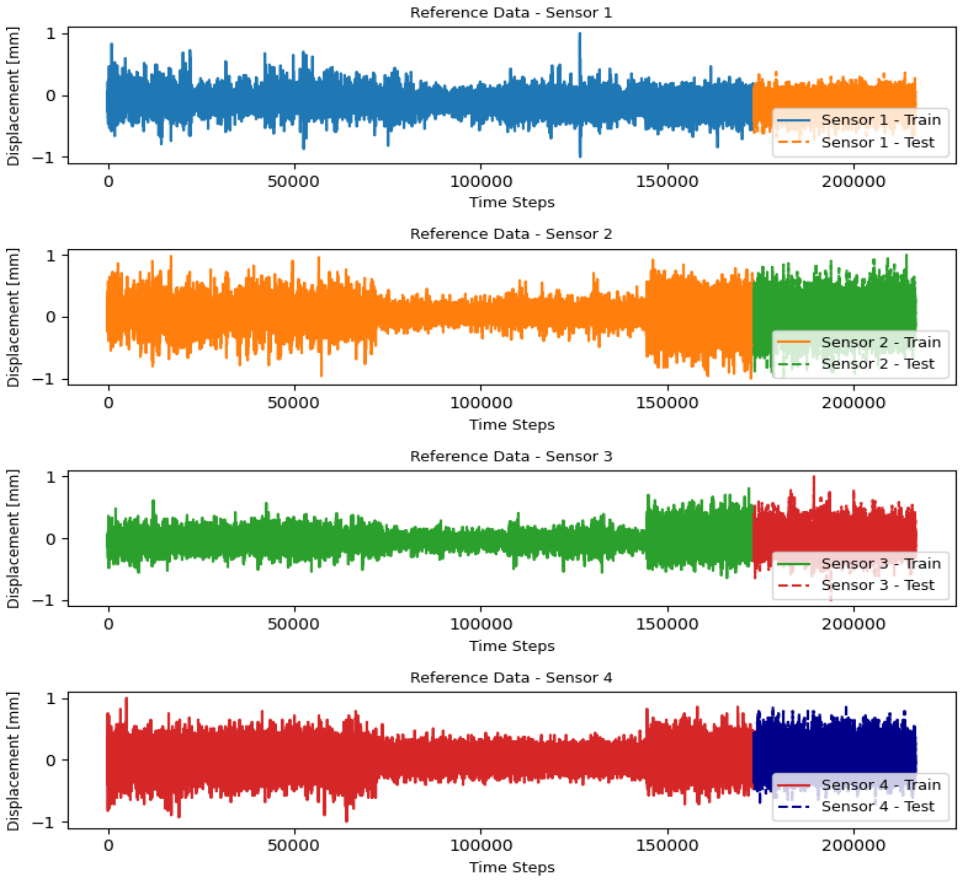


Fig. 5. Displacement-time graphs of the (undamaged) measurements of the first-day a) Sensor-1 horizontal and vertical displacement b) Sensor-2 horizontal and vertical displacement c) Sensor-3 horizontal and vertical displacement d) Sensor-4 horizontal and vertical displacement

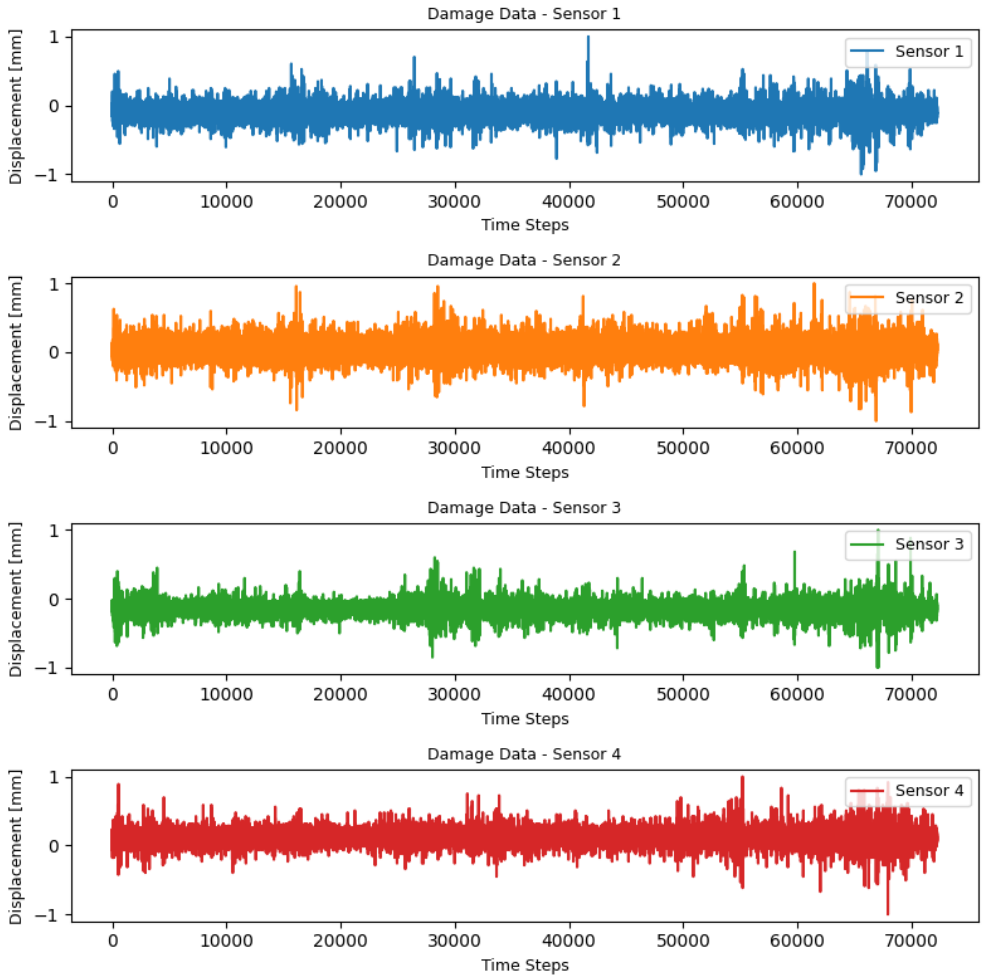


Fig. 6. Displacement-time graphs obtained for Damage-4 after controlled damage a) Sensor-1 horizontal and vertical displacement b) Sensor-2 horizontal and vertical displacement c) Sensor-3 horizontal and vertical displacement d) Sensor-4 horizontal and vertical displacement

5. Results

5.1. CNN model

The CNN model is designed to compress and reconstruct time series data. In the encoder section, sequential 1D convolutional (Conv1D) and max pooling (MaxPooling1D) layers are used to extract and reduce features. The first Conv1D layer applies 64 filters with a kernel size of 3 to capture low-level patterns in the input. This is followed by a MaxPooling1D layer that reduces the temporal resolution by half. The second Conv1D layer uses 128 filters to learn more abstract features, and another pooling layer further compresses the data. By the end of the encoder, the input time series is transformed into a compact latent representation for reconstruction.

The decoder section reconstructs the compressed time series using Conv1DTranspose and UpSampling1D layers. The first transpose convolution layer, with 128 filters, begins expanding the latent representation, followed by an upsampling layer that doubles its size. A second Conv1DTranspose layer with 64 filters further refines the reconstruction, again followed by upsampling. Finally, a third transpose convolution layer with a sigmoid activation restores the time series to its original dimensions, completing the reconstruction process.

The model comprises 99,073 trainable parameters and is optimized using the Mean Absolute Error (MAE) loss function. The Adam optimizer is employed to ensure efficient training. This CNN autoencoder architecture is well-suited for tasks such as feature extraction, anomaly detection, and noise reduction in time series data (Fig. 7). Its structure demonstrates the capability of deep learning models to handle complex temporal patterns, making it a valuable tool for both research and practical applications.

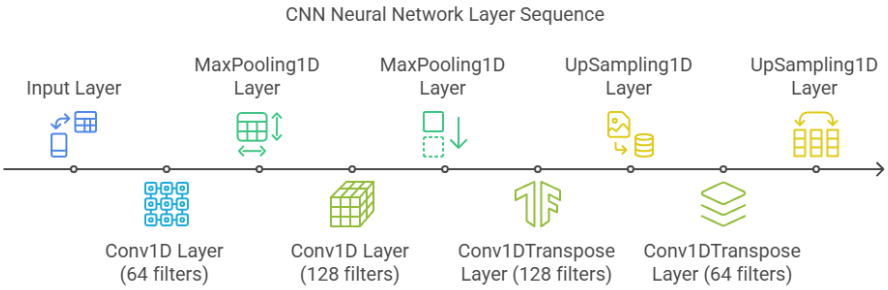


Fig. 7. CNN neural network layer sequence

Fig. 8 shows the training and validation losses of the CNN model. The training loss stabilized at around 0.80, while the validation loss initially decreased from 1.10 to 0.75, then fluctuated between 0.75 and 0.85 during later epochs. The stable training loss suggests that overfitting is not an issue, although the fluctuations in the validation loss might indicate some instability in the model's generalization performance. Overall, the model performs effectively with low loss values. Fig. 9 displays the distribution of MAE losses during training. The losses are predominantly clustered in the 0.4 to 0.6 range, with high loss values being rare. The fact that there are very few samples exceeding the 1.42 threshold indicates that the model performs well on normal data and is effective in anomaly detection. In general, the model has achieved successful learning with low losses. Fig. 10 presents the MAE losses on the CNN model's test data. Most loss values are concentrated below 2, and only a very few samples exceed the threshold of 16.89. This demonstrates that the model successfully reconstructs normal data with low error and is effective in detecting anomalies. Fig. 11 illustrates the CNN model's ability to distinguish between normal and anomalous data. During the time steps 0–250, the model accurately reconstructs normal data with low losses. After the 250th step, it correctly identifies anomalies with high losses (e.g., 46.52 and 237.18). The divergence between the blue and orange lines underscores the model's effective anomaly detection performance.

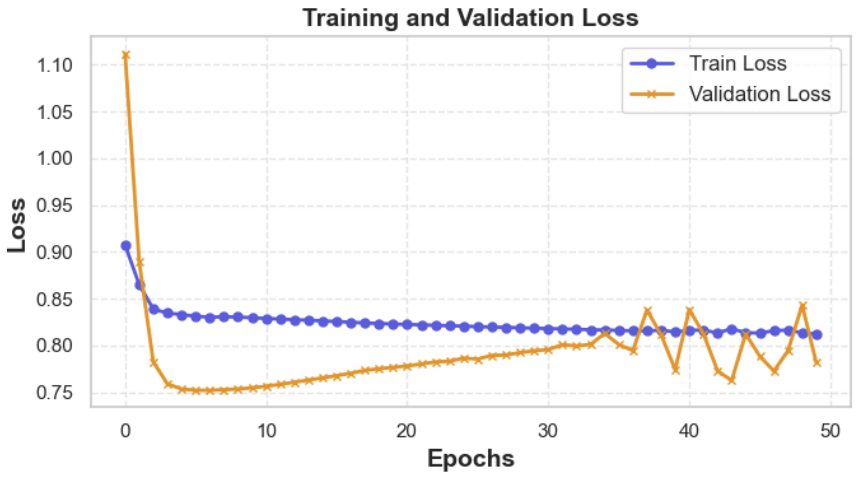


Fig. 8. Training and validation loss graph using CNN model for damage index distribution data

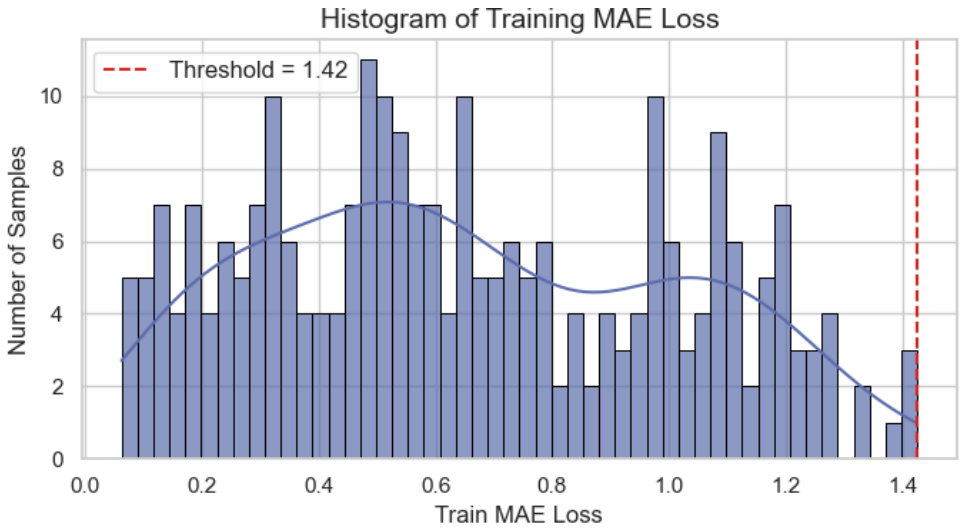


Fig. 9. Histogram of training MAE Loss for CNN model

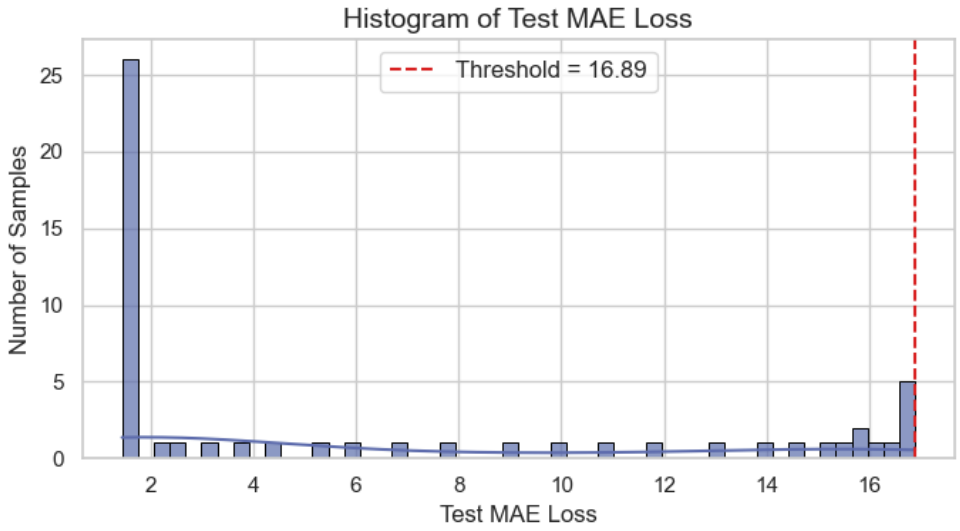


Fig. 10. Histogram of test MAE Loss for CNN model

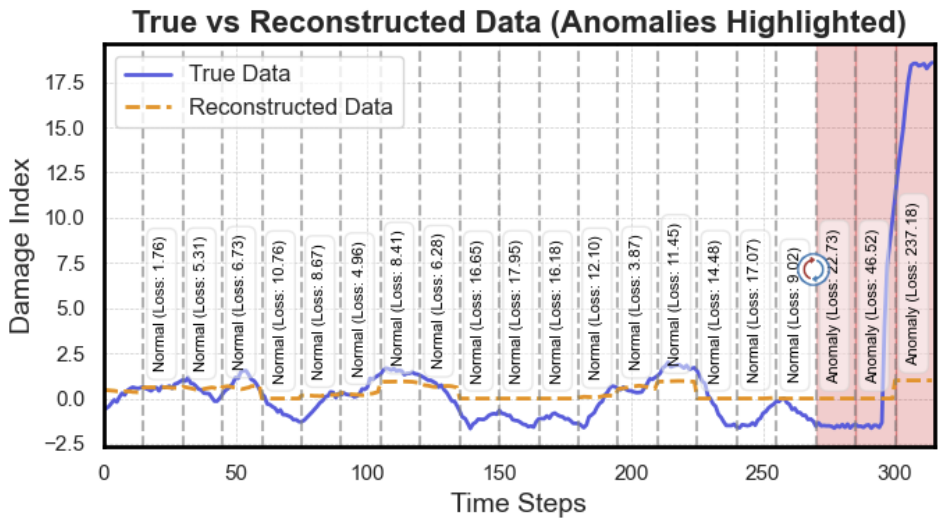


Fig. 11. Anomalies for CNN model

6. Discussion

The proposed video-based structural health monitoring (SHM) system demonstrated considerable promise in assessing the condition of historic masonry minarets through precise, non-contact displacement measurements. By applying advanced optical flow techniques—particularly the pyramidal approach—the system effectively tracked both small and large motions, enabling accurate real-time monitoring of structural dynamics. The conversion of pixel displacements into real-world measurements further confirmed the system's applicability in practical scenarios, especially where traditional sensor-based methods are limited or invasive.

Controlled damage scenarios provided critical insight into the CNN model's performance in anomaly detection. The model exhibited strong spatial feature extraction capabilities, accurately identifying localized damage patterns in the structure. Despite minor fluctuations in validation loss, the CNN maintained stable performance and avoided overfitting, which attests to its reliability for real-world deployment. Its ability to distinguish between normal and anomalous structural conditions underscores the potential of convolutional architectures in video-based SHM applications.

Overall, the findings support the effectiveness of CNNs for spatial anomaly detection in video-based monitoring frameworks. The experimental results validate this approach as a cost-effective, scalable, and non-intrusive tool for continuous structural health assessment, particularly suited for preserving the integrity of historic masonry minarets.

7. Conclusions

This study developed a video-based SHM framework tailored to historic masonry minarets, integrating optical flow techniques with a convolutional neural network (CNN) to achieve accurate, non-invasive anomaly detection. The main findings are as follows:

- Optical flow methods provided reliable, non-contact displacement measurements, enabling structural assessment in areas where direct instrumentation is impractical.
- The CNN model demonstrated strong spatial feature learning capabilities, effectively distinguishing between damaged and undamaged structural states with high accuracy.
- The system's ability to operate under real-time conditions and deliver robust performance under varying damage levels highlights its suitability for long-term structural monitoring.
- The reconstruction-error-based anomaly detection approach enabled the model to generalize from healthy-state data and detect deviations indicative of structural damage.

In conclusion, the proposed video-based SHM system presents a practical and effective approach for monitoring and preserving historic masonry minarets. Its non-invasive nature, combined with the analytical power of CNNs, makes it a valuable contribution to heritage conservation efforts.

Acknowledgments

This study was conducted as part of the research project titled "Video Camera-Based Structural Health Monitoring of Historical Masonry Minarets and Development of a Long-Term Decision-Making Model Based on Deep Learning Method" (Project No: 222M140), supported by the TÜBİTAK 1001 – Scientific and Technological Research Projects Support Program I would like to express my sincere appreciation to the General and Regional Directorates of Foundations for their valuable collaboration and logistical support throughout the research process.

References

Aktürk, F., Aras, F.and Ceribasi, S. (2022). Dynamic response and seismic vulnerabilities of the historic tophane-i amire: dynamic and seismic analysis of a historic masonry building. *J. Comput. Cult. Herit., 15*(3): Article 52.

Baker, S. and Matthews, I. (2004). Lucas-Kanade 20 years on: A unifying framework. *International Journal of* Computer *Vision, 56*, 221-255.

Bouguet, J.-Y. (2001). Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. *Intel Corporation, 5*(1-10), 4.

Dong, C.-Z. and Catbas, F. N. (2021). A review of computer vision-based structural health monitoring at local and *global* levels. *Structural Health Monitoring 20*(2): 692-743.

Farrar, C. R., Dervilis, N. and Worden, K. (2025). The Past, Present and Future of Structural Health Monitoring: An *Overview* of Three Ages. *Strain 61*(1): e12495.

Fioriti, V., Cataldo, A., Colucci, A., Ormando, C., Saitta, F.F., Palumbo, D. and Roselli, I. (2024). Innovative graph-based video processing methodology for collapse early warning of historic *masonry* buildings. *Acta IMEKO, 13*(2).

Ghandil, M., Dabanli, O. and Riahi, H.T. (2021). An enhanced indirect video-based measurement procedure for *dynamic* structural system identification applications. *Measurement 182*: 109759.

Lucas, B. D. and T. Kanade (1981). An iterative image registration technique with an application to stereo vision. In: *IJCAI'81: 7th International Joint Conference on Artificial Intelligence.*

Sabia, D., Demarie, G.V. and Quattrone, A. (2022). Structural health monitoring of historic masonry towers: The Case of the Ghirlandina Tower, Modena. *Geotechnical Engineering for the Preservation of Monuments and Historic Sites III*, CRC Press: 191-201.

Sangirardi, M., Altomare, V., De Santis, S.and De Felice, G. (2022). Detecting damage evolution of masonry structures through computer-vision-based monitoring methods. *Buildings 12*(6): 831.

Shariati, A., Schumacher, T. and Ramanna, N. (2014). Exploration of video-based structural health monitoring techniques, Rutgers University. *Center for Advanced Infrastructure & Transportation.*

Simonyan, K. and A. Zisserman (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv*:1409.1556.

Szegedy, C., Liu, W., Jia, W., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015). Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Xiong, B. (2021). Video-based vibration analysis for structural health monitoring in civil engineering, Université Rennes 1.