

Deep Learning-based Post-earthquake Structural Damage Classification with Small Datasets

Kaveesh Guwanindu ABEYSURIYA*, Mihaela Anca CIUPALA, Meghana MANIKONDA, Kruttika JAMALPURAM, Aishwarya Nitin SONAR, Mhd Saeed SHARIF, Seyed Ali GHORASHI

*School of Architecture, Computing and Engineering
University of East London, United Kingdom*

Alper ILKI

*Faculty of Civil Engineering
Istanbul Technical University, Turkey*

Abstract

This paper introduces a deep learning-based approach for automated classification of local structural element failure modes in post-earthquake buildings using image-based data. Addressing the critical challenge of limited training datasets in the structural/earthquake engineering domain, targeted, domain-informed data augmentation and synthetic data generation techniques are proposed to enhance dataset size and diversity. The model architecture and preprocessing pipeline are explicitly designed to capture damage-sensitive features in images that are essential for informed decision-making on structural integrity of the building, thus extending beyond conventional classification tasks. Dataset enhancement, transfer learning and model regularisation techniques are integrated to ensure alignment of model predictions with expert domain judgement. Achieving 0.93 (93%) accuracy, precision, recall and F1-score, the developed model exhibits robust generalisability without overfitting, demonstrating clear potential for practical deployment in disaster resilience and infrastructure recovery efforts.

Contribution of the Paper: The development of a tailored deep learning-based approach that integrates domain-informed dataset enhancement, transfer learning and targeted regularisation techniques, specifically designed to reliably classify local structural element failure modes from limited, image-based post-earthquake data to support expert-informed structural integrity decisions.

Keywords: deep learning, transfer learning, data augmentation, post-earthquake building damages

IJCVP

ISSN: 2186-1390 (Online)
<http://cennser.org/IJCVP>

Article History:
Received: 12/2/2025
Revised: 15/7/2025
Accepted: 15/8/2025
Published Online: 23/11/2025

© 2012, IJCVP, CNSER. All Rights Reserved

1. INTRODUCTION

Over the past century, numerous major earthquakes have caused partial to complete collapse of building structures worldwide, leading to significant loss of life and severe economic impacts [1]. Investigations carried out after these events have shown that many structures requiring damage inspections can be significantly high [2]. The effects of natural disasters, such as earthquakes, have intensified considerably in recent times [3]. The growing impact of

earthquakes on infrastructure assets, such as buildings, highlights the need for rapid and accurate assessment of their sustained seismic damage. These assessments are essential to enable timely and informed decision-making on occupant safety, understanding structural behaviour during seismic events, and developing effective retrofit strategies. Typically, post-earthquake damage assessments are conducted by qualified engineers in person, on site, in accordance with established post-earthquake damage assessment guidelines and design codes. These assessment processes can be cumbersome and delayed in areas with frequent earthquakes or poor construction practices [4]. Such delays can increase assessment costs and extend the need for temporary housing. Therefore, a

*Corresponding author

Email address: u2467949@uel.ac.uk (Kaveesh Guwanindu ABEYSURIYA)

rapid and accurate assessments that allow for quicker decisions, resulting in retrofitting strategies, that will enable communities to swiftly return to normalcy [5].

Qualitative methods, such as visual inspections, field assessments, and image-based damage detection, are the most common approaches for assessing post-earthquake damage in reinforced concrete (RC) buildings, which is performed by qualified engineers or building inspectors [6]. These approaches visually detect damage indicators, such as cracks, spalling, crushing, buckling of steel reinforcement or residual deformation in structural elements. These damages are then classified manually according to standardised damage assessment guidelines and design codes to assess the building's overall condition. There are several types of classification tasks involved in post-earthquake damage assessments, such as identifying element failure modes, damage features and damage levels, as shown in Figure 1. Among these, classifying element failure modes is the most important for determining the structural integrity of a building. However, relying on manual processes for failure mode classification often results in subjective interpretations, leading to errors or omissions [7]. Moreover, these processes are time-consuming, labour-intensive, and can be compromised by dangerous working conditions, reducing its reliability [8]. Based on these failure mode classifications, engineers use post-inspection structural assessment techniques to simulate the building's response to seismic loads and quantify the induced damage under seismic loads.

There is an urgent need for a rapid and accurate method to classify element failure modes in damaged buildings using image-based data, given the frequent occurrence of earthquakes and the widespread presence of substandard buildings in many disaster-prone areas globally. In response to this demand, deep learning (DL), known for its strong performance in classification and prediction tasks, is increasingly being explored as a promising solution in the structural/earthquake engineering domain. However, it is well established that DL techniques typically require large, balanced datasets to perform classification tasks effectively. This presents a major challenge in structural/earthquake engineering domain, where datasets are inherently small and often unbalanced due to the scarcity of pre-existing repositories and the cost and practical difficulties associated with collecting data. Therefore, this paper investigates methods to enhance dataset size and diversity by evaluating the suitability of various data augmentation and synthetic data generation techniques to simulate realistic damage features in images representing element failure modes - a classification task that has, to date, received limited attention in existing research. Additionally, this paper explores the application of transfer learning and model regularisation techniques to adapt the DL model for improved performance with small, domain-specific datasets, as such in structural/earthquake engineering domain.

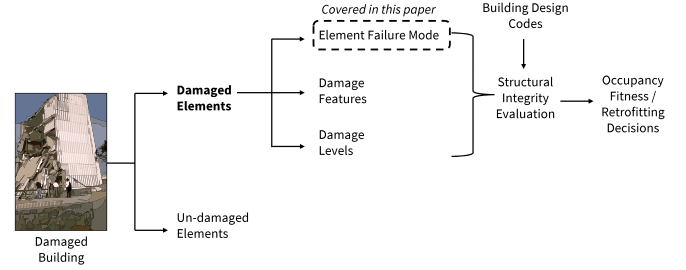


Figure 1: Image-based damage classification tasks contributing to structural integrity evaluations

2. RELATED WORK

The damages in structural local elements, such as columns, beams, masonry walls, and slabs are classified relying on still images, typically captured using handheld cameras, to provide detailed, micro-level visual data of specific structural components, as such damage is often located within the interior of the structure. DL-based approaches are the most widely utilised methods for post-earthquake structural damage classification using image-based data [9]. Among the DL-based approaches, pre-trained convolutional neural network (CNN)-based models are the most employed DL techniques in post-earthquake damage classification, owing to their exceptional capabilities in performing classification and prediction tasks [10]. Therefore, this section critically examines the challenges encountered when applying CNN models to datasets within the structural/earthquake engineering domain. The primary goal is to identify feasible strategies to foster domain generalisation and mitigate overfitting, thereby enhancing the robustness and applicability of pre-trained CNN models within this domain. The network architecture of CNNs is illustrated in Figure 2.

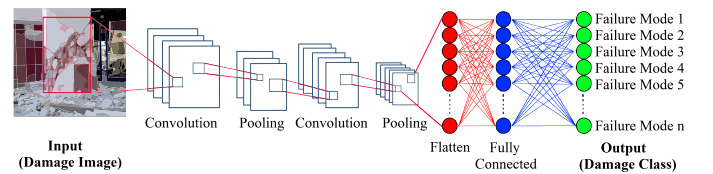


Figure 2: Network architecture of convolutional neural networks

Although widely adopted, CNNs face challenges in the context of image-based damage classification tasks, largely stemming from dataset limitations, modelling issues, and operational constraints. A lack of diversity in datasets in structural/earthquake engineering domain restricts the generalisation of models to varied scenarios, while imbalanced datasets exacerbate issues by skewing predictions towards dominant classes and misclassifying under-represented damage types [11]. Similarly, small

datasets in earthquake /structural engineering domains compared to other disciplines, lead to model overfitting, thereby compromising the model's performance in real-world applications [12]. Moreover, challenges arise in detecting subtle structural damage, such as surface cracks or minor failures, which often go unnoticed due to the limitations of input features or the complexity of damage scenarios [13]. Variability in image quality that is caused by factors, such as lighting, resolution, occlusion and noise, further hinders accurate classification and segmentation [14]. Technical constraints, such as the dependency on manual labelling are labour-intensive and introduce subjectivity, which limits model scalability [15].

Several strategies have been proposed to address these challenges and improve the application of CNNs in damage classification in recent literature. Transfer learning using pre-trained models, can effectively address issues of data scarcity [16]. Transfer learning involves training a new model using pre-trained CNN models, employing either or both feature extraction and fine-tuning methods, as illustrated in Figure 3.

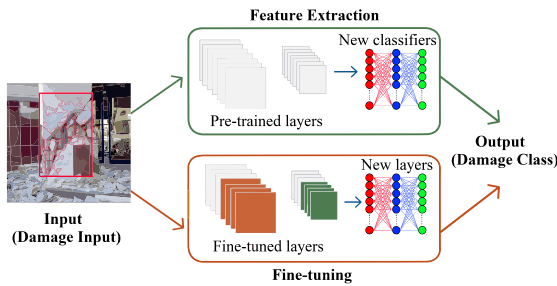


Figure 3: Transfer learning techniques adopted in convolutional neural network models

This approach leverages an existing model and dataset, originally used for a different task, to train a new model on the target dataset. In the feature extraction method, all layers of the pre-trained model are kept frozen, enabling the automatic extraction of generalised features from the new dataset. In contrast, fine-tuning involves unfreezing certain layers, typically the final ones, to adjust the model's hyperparameters, ensuring better alignment with the new dataset. However, the most appropriate pre-trained CNN model, and its performance with feature extraction and/or fine tuning with the target dataset was evaluated before applying the model for the classification/prediction task.

Expanding datasets to include diverse damage geometries and various earthquake scenarios is essential [11]. To achieve dataset expansion, techniques such as data augmentation and synthetic data generation (Figure 4), which enhances variability by simulating real-world conditions are being used. In recent years, researchers rely on augmented datasets due to insufficient

availability of data for training models [17]. However, if augmentation/synthetic data generation techniques are not well-suited to the target dataset, or if there is an over-reliance on these augmented/synthetic data, the applicability of pre-trained CNN models to real-world conditions may be compromised, particularly in novel earthquake scenarios [18]. Therefore, comprehensive comparative study on data augmentation and synthetic data generation techniques, which is essential for identifying the most effective methods for enhancing datasets within the structural/earthquake engineering domain is imperative. This study evaluated commonly used data augmentation and synthetic data generation techniques to determine those most appropriate for the target dataset.

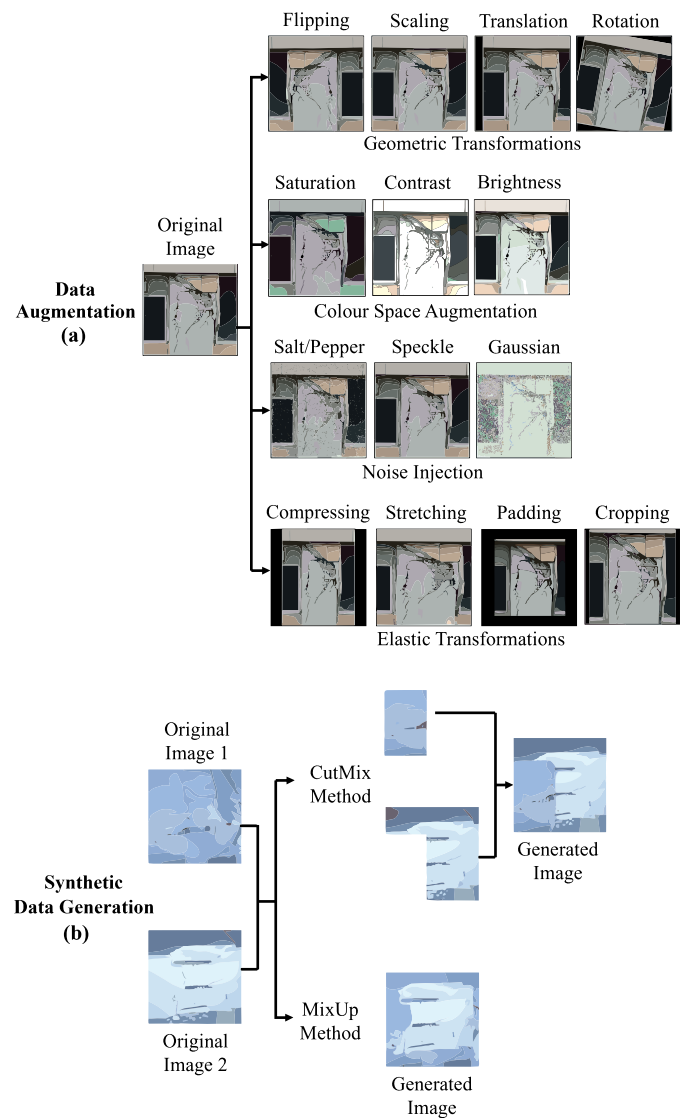


Figure 4: Images generated using data augmentation (a) and synthetic data generation (b) techniques

3. PROPOSED METHOD

To develop a generalisable and non-overfitting damage assessment model, this paper employed a four-step methodology integrating transfer learning, model regularisation with data augmentation and synthetic data generation. The details of this method are presented in the following subsections.

3.1. Dataset Development

Structural element damage is primarily documented using still images captured during on-site building inspections using handheld cameras. These image-based datasets include critical indicators of physical damage in elements, such as crushed regions, crack patterns, and measurements of crack length and width. However, there is a notable lack of publicly available datasets containing diverse and sufficient imagery of post-earthquake damage to structural elements, particularly through web-based repositories. This scarcity is primarily due to the fact that such data are typically collected and retained by individual field investigation teams, limiting open access. A review of existing studies adopting DL techniques for damage classification tasks, revealed that many researchers rely on either self-collected or web-scraped datasets, with total dataset sizes ranging from as few as 50 to around 4,000 images. These dataset sizes are generally considered inadequate for training DL models effectively in classification tasks. To address this limitation, this research first compiled vast majority of available image data from open-source web repositories before applying targeted data augmentation and synthetic data generation techniques to expand and diversify the dataset.

The dataset development process follows a multi-step approach encompassing web scraping, data segregation, data cleaning, quality checks, database structuring, versioning, and documentation. Images are segregated based on the element type of the RC building and the failure mode observed in the damaged element. Columns are the most critical and vulnerable structural elements in RC buildings, as their failure can lead to severe damage or even complete structural collapse. Therefore, this paper focuses specifically on RC columns as the element type and considers its two key failure modes: shear failure and flexural failure, setting the total number of classes for this classification task to two. Future research can extend this work by exploring other structural element types and failure modes. Figure 5 presents a sample of the two-class dataset utilised in this paper.

Table 1 presents the total original image dataset depicting shear failure and flexural failure modes in RC columns, compiled through web scraping the open-source repositories, from which these relevant images were screened and selected for further analysis. This reflects the small

and unbalanced datasets that is characteristic of the structural/earthquake engineering domain.



Figure 5: Reinforced concrete column failure modes in post-earthquake buildings considered in this paper

Table 1: Initial dataset size of reinforced concrete column failure modes considered in this paper

Failure Mode	Total Images
Column Shear Failure	1048
Column Flexural Failure	857

3.2. Dataset Enhancement

To address the limitation of small datasets in the structural and earthquake engineering domain for training DL models, this research explores the enhancement of the previously compiled dataset using data augmentation and synthetic data generation techniques [19]. While similar techniques have been applied in recent studies, there remains a lack of established understanding regarding their suitability for domain-specific datasets [20]. Therefore, it is necessary to evaluate the effectiveness of each enhancement method in replicating the characteristics and diversity of the original data, which is an essential step in developing a model with strong generalisability. The dataset enhancement techniques employed in this research, as outlined in Table 2, were selected through informed decision-making, guided by domain expertise to ensure the realistic simulation of structural damage features. The performance of each dataset enhancement technique is evaluated by training a DL model using each enhancement technique and subsequently validating and testing the model on the original data that is new and unseen to the model before. This approach assesses how effectively each enhancement technique replicates the characteristics of real, domain-specific data, thereby indicating its suitability for mixing into the training dataset of the final classification model.

Table 2: Data augmentation and synthetic data generation techniques considered in this paper

Method	Technique	Parameter
Geometric Transformation	Rotation	5, 10, 15, 20 degrees
	Translation	10, 20, 30 pixels
	Scaling	5%, 10%, 15%
	Flipping	Horizontally
Colour Space Augmentation	Brightness	30%
	Contrast	30%
	Hue/Saturation	60 degrees, 30% saturation factor
Noise Injection	Gaussian	Mild - Mean 0, Standard Deviation 10
	Speckle	Mild - 0.1 probability from Gaussian
	Salt-and-pepper	Mild - 0.02 probability
Elastic Transformation	Cropping	10%
	Padding	5%
	Stretching	5%
	Compressing	5%
Synthetic Data Generation	MixUp	Combines two images and their labels by taking a weighted average. This creates interpolated training samples
	CutMix	Cuts a portion of one image and pastes it onto another, mixing labels proportionally
	CutOut	Randomly masks out parts of the image to force the model to focus on the surrounding context
	GridMask	Masks out parts of the image in a grid-like pattern to diversify learned features
	GridMix	Mix and recombine parts of the various images in a grid-like pattern to ensure statistical utility in a new image

3.3. Transfer Learning

3.3.1. Pre-trained Models with Feature Extraction

Various pre-trained models that have been utilised for feature extraction to train models with new data from the structural/earthquake engineering domain were identified in literature. To select the most appropriate model for the dataset of this research, the following models have been chosen and compared based on their widespread use and demonstrated positive performance in the domain of image-based damage assessment: VGGNet (VGG16 and VGG19) [21], GoogleNet (Inception) [22], Xception [23], MobileNet [24], DenseNet (DenseNet121, DenseNet169, DenseNet201) [25], ResNet (ResNet50, ResNet101 and ResNet152) [26], EfficientNet-B0 [27] and AlexNet [28]. All the layers in these models have been kept frozen, enabling the models to extract features from the new target dataset. As these pre-trained models were originally developed using datasets from other domains, it is essential to evaluate their performance in learning features from new target datasets.

3.3.2. Pre-trained Models with Fine Tuning

The pre-trained models exceeding the overall accuracy, precision, recall and F1 score threshold of 0.70 and meeting the criteria for overfitting and generalisation checks are selected for the fine-tuning stage. This step narrows the pre-trained models listed above, ensuring that only the most appropriate models are evaluated during the fine-tuning. At this point, the selected pre-trained models, which have already extracted features during feature extraction stage, have a selected number of deeper (top)

layers unfrozen. This approach aims to balance the retention of pre-learned generic features with the acquisition of dataset-specific knowledge, ultimately enhancing the model's generalisability.

3.3.3. Model Regularisation Parameters

The pre-trained model that demonstrates the best performance after the feature extraction and fine-tuning stages is subjected to a comparative analysis of regularisation parameters, such as the loss function, regularisation factor, dropout rate and learning rate. Additionally, model parameters, such as the number of epochs and the number of unfrozen layers are evaluated to optimise performance further.

3.4. Model Evaluation

Evaluating the performance of the DL model is essential for assessing how effectively the model generalises to new and unseen data that will be provided during field investigations. A confusion matrix serves as a critical tool that provides a comprehensive breakdown of the model's predictions by comparing them to the actual outcomes [29]. The confusion matrix consists of four key elements:

- True Positives (TP): The model correctly predicted the positive class
- True Negatives (TN): The model correctly predicted the negative class
- False Positives (FP): The model incorrectly predicted the positive class

- False Negatives (FN): The model incorrectly predicted the negative class

A confusion matrix helps derive several performance metrics, such as accuracy, precision, recall, and F1 score, which provide deeper insights into model performance [29].

- Accuracy measures the proportion of correct predictions out of the total, giving an overall sense of the model's performance. However, it can be misleading if the dataset is imbalanced, with significantly more instances of one class than the other

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

- Precision, or positive predictive value, is the ratio of correctly predicted positives to all predicted positives. It is important when minimizing false positives is critical

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

- Recall measures the model's ability to correctly identify all positive instances, making it essential in cases where missing positive instances (false negatives) is costly

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

- F1 score, the harmonic mean of precision and recall, provides a balanced metric, especially useful when class distribution is uneven. It balances the trade-off between precision and recall, offering a comprehensive view of model performance

$$F_1\text{-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

To evaluate whether the developed DL model is overfitting or capable of generalising to new and unseen data, the following model evaluation techniques, such as model overfitting and generalisability checks are employed [30]:

- Check for model overfitting using training accuracy (TRA) and validation accuracy (VLA) for both feature extraction and fine-tuning phases [30]:

$$\text{Training to Validation Gap} = TRA - VLA \quad (5)$$

If this gap is less than, 0.05 the model is considered safe from overfitting.

- Check for model generalisability using fine-tuning validation accuracy (VLA) and overall model test accuracy (TSA) [30]:

$$\text{Accuracy Gap} = VLA - TSA \quad (6)$$

If this gap is less than, 0.03 the model is considered to be well generalisable.

- Check for model generalisability using fine-tuning validation loss (VLL) and overall model test loss (TSL) [30]:

$$\text{Loss Gap} = VLL - TSL \quad (7)$$

If this gap is less than, 0.001 the model is considered to be well generalisable.

4. RESULTS

4.1. Comparison of Pre-trained Models with Feature Extraction

The performance indicators (accuracy, precision, recall and F1 score) for the pre-trained models that utilised the feature extraction transfer learning technique is presented in Figure 6. The top-performing models from this evaluation, with their respective accuracy, precision, recall and F1 score values in brackets, include DenseNet169 (0.84, 0.84, 0.84, 0.84), MobileNet (0.83, 0.83, 0.83, 0.83), DenseNet201 (0.81, 0.82, 0.81, 0.81), GoogleNet (0.80, 0.80, 0.80, 0.80), DenseNet121 (0.78, 0.79, 0.78, 0.78), Xception (0.77, 0.77, 0.77, 0.77), and VGG16 (0.77, 0.77, 0.77, 0.77). Only these models were subsequently selected for the next stage of the study, where the model is fine-tuned. It is noteworthy that MobileNet demonstrated the shortest runtime, completing the task in 32 minutes, while the VGG16 model had the longest runtime, with an elapsed time of 4 hours and 23 minutes. In this study, the models were executed in Python using a cloud-based graphical processing unit (GPU), specifically an NVIDIA L4 tensor Core GPU with a 22.5 GB random access memory (RAM) available on Google Colab Pro platform. The dotted line corresponds to the 0.70 threshold accuracy, precision, recall and F1 score was considered for selecting the models for fine-tuning.

4.2. Comparison of Pre-trained Models with Fine Tuning

Similarly, the performance indicators (accuracy, precision, recall and F1 score) for the selected models that were utilised in fine-tuning stage having achieved over 0.70 benchmarking accuracy, precision, recall and F1 score in the feature extraction stage is presented in Figure 7. The top-performing model was MobileNet (accuracy, precision, recall and F1 score: 0.88) which was also the model that demonstrated the shortest runtime, completing the task in 43 minutes, having improved its performance since feature extraction stage.

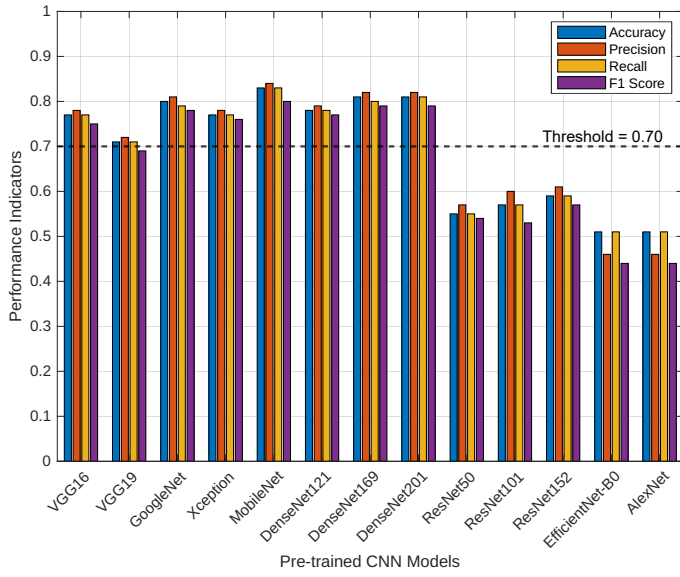


Figure 6: Accuracy, Precision, Recall and F1 Score of pre-trained models with feature extraction

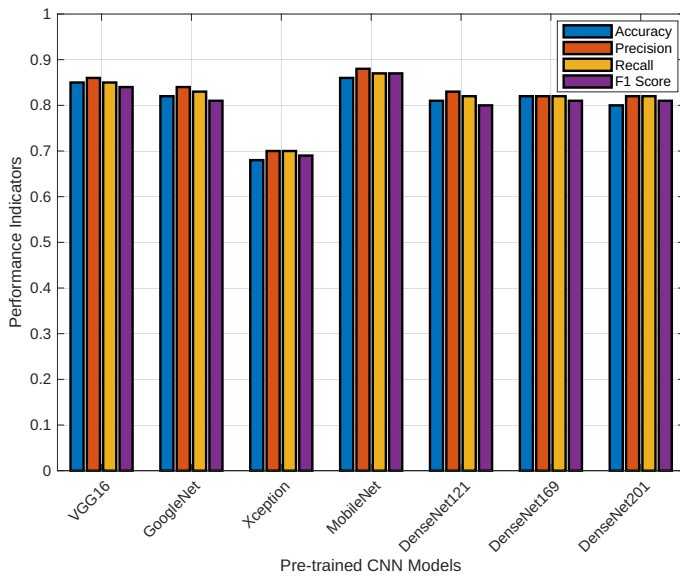


Figure 7: Accuracy, Precision, Recall and F1 Score of pre-trained models with feature extraction and fine-tuning

4.3. Comparison of Model Regularisation Parameters

The regularisation parameters for the MobileNet model were compared to identify the optimal loss function regularisation factor, dropout rate, and learning rate. Additionally, the optimal model parameters, including the number of epochs and the number of unfrozen layers, were evaluated to achieve model generalisation and minimise overfitting. Table 3 summarises the optimum parameters identified through this comparative study.

4.4. Comparison of Dataset Enhancement Techniques

The performance of the modified MobileNet model, trained using data generated from each augmentation or synthetic data generation technique and validated and tested with the data from the original dataset, was evaluated. The performance indicators (accuracy, precision, recall and F1 score) for each augmentation technique are presented in Figure 8. The dotted line corresponds to the 0.70 threshold accuracy, precision, recall and F1 score was considered for selecting the dataset enhancement techniques for the final model. This analysis reveals that traditional data augmentation methods, such as geometric transformations, elastic transformations, colour space augmentation, and noise injection, performed better in training the model due to their close similarity to the original data. Synthetic data generation techniques, particularly, CutMix and MixUp, demonstrated reasonably strong performance and proved useful in achieving greater dataset variation. Consequently, a combination of these techniques (refer Table 4) was employed to develop the final dataset used for training the final model.

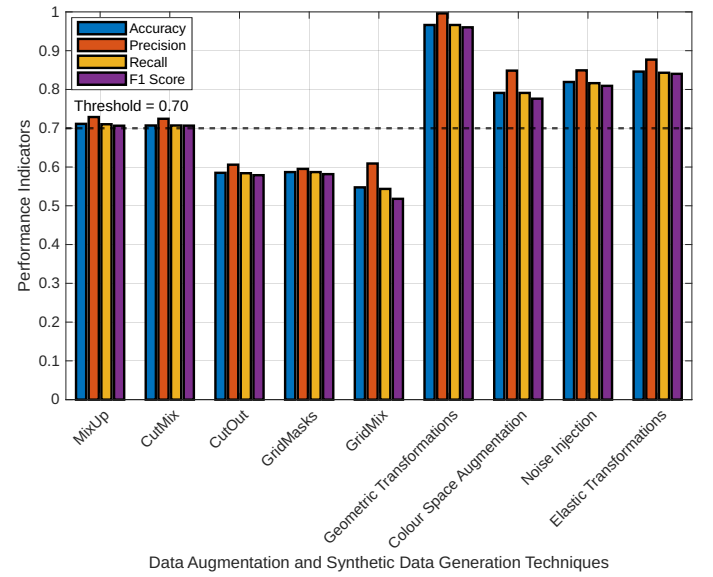


Figure 8: Accuracy, Precision, Recall and F1 Score of MobileNet trained with data augmentation and synthetic data generation techniques

4.5. Final Model

A final model that combined the original data in Table 1 and the augmented data, resulting in a total dataset of 6156 images for shear failure and 6105 images for flexural failure, is tabulated in Table 5. The final model's confusion matrix is presented in Figure 9.

Table 3: Optimum model regularisation parameters and model parameters for MobileNet pre-trained model

Type	Parameter	Optimum Value
Model Parameters	Image Size	224x224
	Batch Size	32
	Epochs	25
	Total Layers in Model	28
	Unfreezing Layers	10
Regularisation Parameters	Learning Rate	0.00001
	Loss Function (L2) Regularisation Factor	0.02
	Dropout Rate	0.6
	Activation Function	Softmax (dense)
	Optimiser	Adam
	Learning Rate Scheduling	λ epoch: $1e-5 * 10^{(-epoch/10)}$
	Early Stopping Patience	2
	Batch Normalisation	Yes

Table 4: Final dataset combining data augmentation and synthetic data generation techniques considered in this paper

Method	Technique	Shear Failure	Flexural Failure
Geometric Transformations	Rotation	802	857
	Translation	802	857
	Scaling	802	857
	Flipping	802	857
Colour Space Augmentation	Brightness	100	100
	Contrast	100	100
	Hue/Saturation	100	100
Noise Injection	Gaussian	200	200
	Speckle	200	200
	Salt-and-pepper	200	200
Elastic Transformations	Cropping	200	200
	Padding	200	200
	Stretching	200	200
	Compressing	200	200
Synthetic Data	MixUp	100	100
	CutMix	100	100
Original Images		1048	857
Total		6156	6105

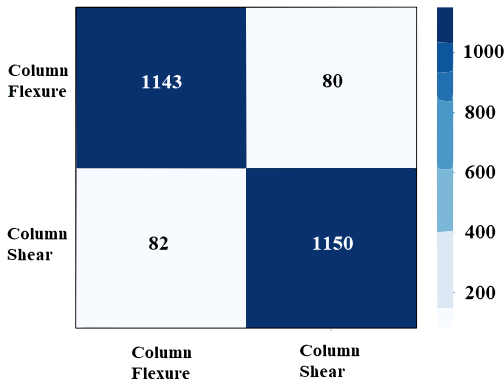


Figure 9: Confusion matrix of final model

As indicated in Table 6, this model attained scores rounding up to 0.93 (93%) for accuracy, precision,

recall, and F1 score. Notably, achieving identical values across these four-performance metrics confirms the dataset's balanced nature, ensuring that false classifications/predictions in one class do not disproportionately affect others. Furthermore, the results from overfitting and generalisation checks (Table 6) show that the model reliably predicts outcomes for new and unseen data.

4.6. Future Directions

Based on the results presented in this paper, the following future research directions are proposed:

- Advancing on synthetic and simulated data generation via generative adversarial networks (GAN) and/or 3D modelling tools to further enrich model training;
- Incorporating explainability techniques, such as Grad-CAM to highlight visual damage features in images to support classification decision making;

Table 5: Final dataset size of the RC column failure modes considered in this paper

Failure Mode	Training (70%)	Validation (10%)	Testing (20%)	Total
Shear Failure	4308	616	1232	6156
Flexural Failure	4270	612	1223	6105

Table 6: Final model performance results

Criteria	Matrix / Check	Shear Failure	Flexural Failure	Limit
Performance	Accuracy	0.9340	0.9340	N/A
	Precision	0.9350	0.9331	N/A
	Recall	0.9334	0.9346	N/A
	F1 Score	0.9342	0.9338	N/A
Overfitting	Training accuracy and validation accuracy difference (feature extraction)		0.0002	0.05
	Training accuracy and validation accuracy difference (fine-tuning)		0.0100	0.05
Generalisation	Fine-tuning validation accuracy and overall model test accuracy gap		0	0.03
	Fine-tuning validation loss and overall model test loss gap		0.0002	0.001

- Real-world validation of the developed model by deploying in field assessments in disaster prone regions;
- Exploring advanced learning paradigms, such as few-shot and meta-learning approaches to address data scarcity;
- Integration of multi-modal data, such as structural engineering metadata and vibration-based data to improve model robustness.

5. CONCLUSIONS

This paper presents a novel deep learning-based approach specifically tailored for the automated classification of local structural element failure modes in post-earthquake reinforced concrete buildings using image-based data. Given that traditional visual assessment methods are time-consuming, subjective and potentially hazardous, this research directly addresses the critical limitation of small and unbalanced datasets typical in structural/earthquake engineering domain. Through targeted, domain-informed data augmentation and synthetic data generation techniques, this research effectively enhanced dataset diversity and size, significantly mitigating overfitting and improving model generalisation capabilities. By conducting a comparative evaluation of pre-trained convolutional neural networks, MobileNet was identified as the most suitable backbone architecture for the local element failure mode classification task, achieving notable improvements from an initial accuracy of 0.83 (83%) to 0.93 (93%) across all performance indicators used in this paper: accuracy, precision, recall and F1 score. The tailored model architecture and preprocessing pipeline

ensured the extraction of damage-sensitive visual features directly relevant for informed structural integrity decisions, distinguishing this work from standard classification approaches. Moreover, the key future research directions proposed in this paper include, enhancing synthetic and simulated data generation through generative adversarial networks (GAN) and 3D modelling tools to improve model training, integrating explainability techniques, such as Grad-CAM to highlight visual damage features and support classification decisions, and validating the developed model in real-world scenarios through field deployment in disaster-prone regions. These findings establish transferable knowledge by effectively integrating advanced AI methodologies with traditional structural/earthquake engineering practices, together with the proposed future research directions offer practical pathways to enhance urban resilience and strengthen global disaster preparedness.

ACKNOWLEDGEMENT

The financial contribution of the University of East London, United Kingdom, the Daiwa Anglo-Japanese Foundation, United Kingdom, and the 2024 Earthquake Engineering Field Investigation Team (EEFIT) Research Grant Scheme, the Institution of Structural Engineers, United Kingdom is greatly appreciated.

References

- [1] Y. Alberto, O. M, H. Kyokawa, T. Kiyota, I. Towhata, Reconnaissance of the 2017 Puebla, Mexico earthquake., Soils Found 58 (2018) 1073–1092.
- [2] M. Kazama, T. Noda, Damage statistics (summary of the 2011 off the Pacific coast of Tohoku earthquake damage), Soils Found 52 (2012) 780–792.

- [3] B. J. J. R.-V. L. F. P. R. M. G, C. H. C. G. M, Development of seismic vulnerability assessment methodologies over the past 30 years, *ISET Journal of Earthquake Technology* 43 (2006) 75–104.
- [4] T. Gurbuz, A. Cengiz, S. Kolemenoglu, C. Demir, A. Ilki, Damages and failures of structures in İzmir (turkey) during the october 30, 2020 aegean sea earthquake, *Journal of Earthquake Engineering* 27 (2023) 1565–1606. doi:10.1080/13632469.2022.2086186.
- [5] T. Rossetto, A. Elnashai, Derivation of vulnerability functions for european-type rc structures based on observational data, *Engineering Structures* 25 (2003) 1241–1263. doi:10.1016/S0141-0296(03)00060-9.
- [6] M. Gatti, Structural health monitoring of an operational bridge: A case study, *Engineering Structures* 195 (2019) 200–209. doi:https://doi.org/10.1016/j.engstruct.2019.05.102. URL <https://www.sciencedirect.com/science/article/pii/S0141029618336563>
- [7] B. A. Graybeal, B. M. Phares, D. D. Rolander, M. Moore, G. Washer, Visual inspection of highway bridges, *Journal of Nondestructive Evaluation* 21 (2002) 67–83. doi:10.1023/A:1022508121821.
- [8] H. Kim, E. Ahn, M. Shin, S.-H. Sim, Crack and noncrack classification from concrete surface images using machine learning, *Structural Health Monitoring* 18 (2019) 725–738. doi:10.1177/1475921718768747.
- [9] B. Gultekin, G. Dogan, A novel approach to discriminate between structural and non-structural post-earthquake damage in rc structures, *Advances in Civil Engineering* 2024. doi:10.1155/2024/6027701.
- [10] S. Katsigiannis, S. Seyedzadeh, A. Agapiou, N. Ramzan, Deep learning for crack detection on masonry façades using limited data and transfer learning, *Journal of Building Engineering* 76. doi:10.1016/j.job.2023.107105.
- [11] S. Azhari, A. Mahmoodi, A. Samavi, M. Hamidia, Multi-feature driven seismic damage state identification for reinforced concrete shear walls using computer vision and machine learning, *Advances in Engineering Software* 199. doi:10.1016/j.advengsoft.2024.103796.
- [12] M. S. Barkhordari, D. J. Armaghani, P. G. Asteris, Structural damage identification using ensemble deep convolutional neural network models, *CMES - Computer Modeling in Engineering and Sciences* 134 (2023) 835–855. doi:10.32604/cm.2022.020840.
- [13] M. Y. Cheng, M. N. Sholeh, A. Kwek, Computer vision-based post-earthquake inspections for building safety assessment, *Journal of Building Engineering* 94. doi:10.1016/j.job.2024.109909.
- [14] M. Yilmaz, G. Dogan, M. Arslan, A. Ilki, Categorization of post-earthquake damages in rc structural elements with deep learning approach, *Journal of Earthquake Engineering* 28 (2024) 2620–2651. doi:10.1080/13632469.2024.2302033.
- [15] M.-Y. Cheng, R. R. Khasani, R. J. Citra, Image-based preliminary emergency assessment of damaged buildings after earthquake: Taiwan case studies, *Engineering Applications of Artificial Intelligence* 126 (2023) 107164. doi:10.1016/j.engappai.2023.107164.
- [16] J. Liu, Y. Luo, S. Chen, J. Wu, Y. Wang, Bdhe-net: A novel building damage heterogeneity enhancement network for accurate and efficient post-earthquake assessment using aerial and remote sensing data, *Applied Sciences (Switzerland)* 14. doi:10.3390/app14103964.
- [17] Z. Bai, T. Liu, D. Zou, M. Zhang, Q. Hu, A. Zhou, Y. Li, Multi-scale image-based damage recognition and assessment for reinforced concrete structures in post-earthquake emergency response, *Engineering Structures* 314. doi:10.1016/j.engstruct.2024.118402.
- [18] J. Ye, H. Yu, G. Liu, J. Zhou, J. Shu, Component identification and depth estimation for structural images based on multi-scale task interaction network, *Buildings* 14. doi:10.3390/buildings14040983.
- [19] J. Shijie, W. Ping, J. Peiyi, H. Siping, Research on data augmentation for image classification based on convolution neural networks, in: 2017 Chinese Automation Congress (CAC), 2017, pp. 4165–4170. doi:10.1109/CAC.2017.8243510.
- [20] A. Mikołajczyk, M. Grochowski, Data augmentation for improving deep learning in image classification problem, in: 2018 International Interdisciplinary PhD Workshop (IIPhDW), 2018, pp. 117–122. doi:10.1109/IIPhDW.2018.8388338.
- [21] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: 3rd International Conference on Learning Representations (ICLR 2015), Computational and Biological Learning Society, 2015, pp. 1–14.
- [22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014) 1–9. URL <https://api.semanticscholar.org/CorpusID:206592484>
- [23] F. Chollet, Xception: Deep learning with depthwise separable convolutions, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 1800–1807. URL <https://api.semanticscholar.org/CorpusID:2375110>
- [24] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, *ArXiv abs/1704.04861*. URL <https://api.semanticscholar.org/CorpusID:12670695>
- [25] G. Huang, Z. Liu, L. V. D. Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2261–2269. doi:10.1109/CVPR.2017.243.
- [26] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 770–778. URL <https://api.semanticscholar.org/CorpusID:206594692>
- [27] M. Tan, Q. V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, *ArXiv abs/1905.11946*. URL <https://api.semanticscholar.org/CorpusID:167217261>
- [28] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, *Communications of the ACM* 60 (2017) 84–90. doi:10.1145/3065386.
- [29] D. Powers, Evaluation: From precision, recall and f-factor to roc, informedness, markedness correlation, *Mach. Learn. Technol.* 2.
- [30] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, 2nd Edition, Vol. 1, MIT Press, 2016.