Vol.9 Issue 1 Jan 2026

Multimedia Research

Received 12 August, Revised 10 October, Accepted 3 January

Detection of Varieties of Cracks in Concrete Structures Using Selected Machine-Learning Techniques

Michael Ikuomola Akinbo
Department of Civil and Environmental Engineering, Federal University of Technology, Akure. Nigeria.
ikumichaelakimbo@gmail.com

Oluwafemi Omotayo
Department of Civil and Environmental Engineering, Federal University of Technology, Akure. Nigeria.
ooomotayo@futa.edu.ng

David John Omeiza
Department of Computer Engineering, Federal University of Technology, Akure. Nigeria.
davischoice@gmail.com

Abstract: This paper presents a multimedia image analysis framework for detecting cracks in concrete structures using deep learning models. The motivation for this study stems from the increasing need for automated structural health monitoring to reduce manual inspection costs and enhance safety. Crack images are treated as structural multimedia data, positioning the study within computer vision and content-based image processing. This framing allows for leveraging advanced multimedia processing techniques to extract meaningful features from complex structural imagery. Two architectures, ResNet50 and Vision Transformer (ViT), were applied to a curated dataset of 10,000 crack images categorized into hairline and structural cracks. The dataset was meticulously annotated and balanced to ensure model reliability across different crack types and sizes. Preprocessing and augmentation were implemented as multimedia content transformation steps to enhance robustness under varying imaging conditions. Augmentation techniques included rotation, scaling, and contrast adjustments to simulate real-world inspection scenarios. Experimental results show strong performance, with ResNet50 achieving 99.80% testing accuracy and ViT 99.40%. These results indicate that both CNN-based and transformer-based approaches can effectively capture intricate patterns in structural defects. Beyond structural health monitoring, the approach contributes to multimedia analytics by demonstrating how transformer-based and residual networks can generalize to domains including remote sensing, industrial defect recognition, and automated surveillance. Moreover, the study provides insights into model interpretability, showing which image regions most strongly influence predictions, thus aiding trust in automated systems. The findings support integration into UAV inspections and robotic monitoring systems, advancing intelligent multimedia-based structural health assessment.

Keywords: Concrete Crack Detection, Resnet50 Model, Vision Transformer (VIT) Model, Deep Learning, Machine Learning, Image Classification, Computer Vision, Multimedia Image Analysis.

Nomenclature

Abbreviation	Description
ANN	Artificial Neural Network
AOA	Adam Optimizer Algorithm
CNN	Convolutional Neural Network
FC	Fully Connected layer
ML	Machine Learning
NDT	Non-Destructive Testing
NLP	Natural Language Processing
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
SVM	Support Vector Machine
ViT	Vision Transformer

1. Introduction

Civil infrastructure monitoring is increasingly framed as a multimedia image analysis problem, where structural cracks are interpreted as visual data requiring advanced computational analysis. With the rapid growth of smart cities and sensor networks, integrating data analytics and intelligence into civil infrastructure has become crucial for proactive maintenance and safety assurance. Traditional non-destructive testing methods, such as ultrasonic or electromagnetic testing, are limited by localized inspection and high cost. These conventional approaches also suffer from subjectivity and inconsistency

https://doi.org/10.46253/j.mr.v9i1.a2

due to reliance on manual interpretation. In contrast, multimedia-driven approaches leverage computer vision and deep learning to detect, classify, and evaluate crack patterns from large-scale visual datasets. By transforming structural inspection into a data-driven multimedia problem, analysts can extract actionable insights from diverse imaging modalities, including UAV imagery, mobile photography, and drone videos.

To ensure the safety and functionality of building structures, it is imperative to have an effective crack detection system in place [1]. Despite the promising performance of deep learning in structural health monitoring, existing studies often focus on small datasets or single crack types, limiting generalizability across real-world scenarios. By embracing this latest technology, we can propel the construction industry towards automated and intelligent civil engineering structures [2]. Additionally, the lack of integration between multimedia analytics and intelligent robotics platforms poses a challenge for scalable and autonomous monitoring solutions.

By reframing crack detection as a multimedia problem, this study highlights opportunities for cross-domain applications and intelligent automation in structural health monitoring. The primary objective of this study is to develop a robust multimedia-based framework capable of accurately detecting and classifying various crack types using deep learning architectures. Secondary objectives include evaluating the generalizability of the framework to different imaging conditions and exploring its potential integration with UAV- and robot-based inspection systems.

The study addresses the following research questions: (1) Can multimedia-based deep learning models reliably distinguish between hairline and structural cracks across diverse datasets? (2) How do CNN-based and transformer-based architectures compare in terms of accuracy and robustness for crack detection? It is hypothesized that combining image preprocessing, augmentation, and advanced deep learning models will significantly enhance detection performance and system reliability. This research contributes to both the fields of data analytics and civil engineering by demonstrating how intelligent multimedia systems can automate structural monitoring, reduce inspection costs, and improve safety outcomes. The study is limited to visual crack inspection from static images and does not cover embedded sensor data or dynamic structural responses, though these are suggested directions for future research.

2. Literature Review

Crack detection in structures is essential for ensuring their safety, integrity, and longevity. Various methods and technologies have been adopted, developed, and refined over the years to accurately detect cracks in different types of civil infrastructures. This literature review explores some of the prominent techniques and advancements in crack detection, such as Visual Inspection, NDT, image processing techniques and Machine learning. Recent developments in data analytics and intelligent multimedia processing have further enabled automated and scalable crack detection, reducing reliance on manual inspection.

2.1 Visual Inspection

Visual inspection is the simplest mode of detecting cracks compared to other modes. In this method, the structure is looked at directly by the naked eye to find a change in the pattern caused by a crack. According to researches analysis, emphasis was placed on crack width as a critical parameter in assessing the structural integrity of concrete elements. However, visual inspection is highly subjective, and its reliability depends on the inspector's experience and the visibility conditions of the structure. Visual inspection involves measuring and evaluating crack width, length, pattern, and location to determine whether they are within acceptable limits defined by relevant codes and standards. Despite its simplicity, visual inspection remains a standard initial assessment method, often used to guide further non-destructive testing or image-based analysis [16].

2.2 NDT

NDT techniques such as ultrasonic testing, radiographic testing, and magnetic particle testing are majorly classified as traditional methods of crack detection in structures like that of Visual inspection. But they offer more reliable results compared to visual inspections. According to research made by [3], these techniques can be used to denote internal crack detection without causing damage to the structure. Ultrasonic testing, in particular, has been widely used due to its ability to perceive both surface and subsurface cracks [4]. Nevertheless, NDT methods require specialized equipment and trained personnel, limiting their scalability for large infrastructures.] [Recent studies suggest integrating NDT with image processing and AI techniques to enhance detection accuracy while reducing inspection time [17].

2.3 Image Processing Techniques

With advancements in digital imaging and computer vision, image processing techniques have gained reputability in crack detection. These is archived by capturing high-resolution images of the structure and using algorithms to identify cracks. Edge detection algorithms, such as the Canny edge detector, are mostly used to highlight crack features [5]. Preprocessing steps like noise reduction, contrast enhancement, and morphological operations are often applied to improve the clarity of cracks before algorithmic analysis. Moreover, combining image processing with machine learning models has shown improved robustness in detecting cracks under varying lighting and surface conditions [18][21].

2.4 Machine Learning Approaches

Machine learning is a branch of artificial intelligence that aimed at developing algorithms and statistical models that enable computer systems to learn and improve from data without being explicitly programmed [6] and [7]. ML algorithms have revolutionized multimedia image detection by automating the analysis process and improving accuracy. CNNs have been particularly effective in image-based crack detection, learning from large datasets to distinguish cracks from other features [8][9][15]. Deep learning models can automatically learn hierarchical features from raw image data, reducing the need for manual feature engineering. SVM and ANN are also used for classifying and predicting crack presence. The basic premise of machine learning is to build algorithms that can receive input multimedia data and use statistical analysis to predict an output while updating outputs as new multimedia data becomes available [19]. The adaptability of ML models allows them to handle diverse imaging conditions, such as different textures, lighting, and noise levels.

2.5 Present Research Status

Numerous studies have highlighted that CNN models are widely applied in multimedia image detection because of their ability to automatically extract features directly from raw image data. Research findings indicate that CNNs can achieve high accuracy in identifying different types of cracks in concrete structures [16]. For example, [10] employed a deep CNN to detect cracks in concrete bridges, reporting substantial improvements in accuracy compared to conventional approaches. Similarly, [11] demonstrated the effectiveness of transfer learning by fine-tuning a pre-trained VGG16 model on ImageNet for crack detection, achieving high accuracy despite limited training samples. Hybrid approaches, which integrate machine learning algorithms with traditional image processing methods to improve detection accuracy, were also explored in [9]. Such approaches often include preprocessing techniques such as edge detection and filtering to enhance image quality before feeding it into machine learning models. In fact, [9] combined traditional image processing with CNN-based methods for detecting cracks in concrete tunnels, achieving strong performance under challenging lighting variations. Despite these advances, challenges remain in generalizing models across different infrastructures and imaging conditions [17].

With the progress of computer vision, advanced deep learning models such as ViT and ResNet50 have significantly improved automated multimedia image detection. Several works have confirmed the effectiveness of ResNet50 in recognizing and classifying cracks across different conditions, including varying lighting and texture, consistently outperforming classical methods such as edge detection and basic CNNs. Specifically, [12] applied ResNet50 for automatic crack detection in concrete surfaces, obtaining high precision and recall by fine-tuning the model on an extensive dataset of crack and non-crack images. The results also showed that it surpassed traditional machine learning techniques in noisy environments. In addition, [13] utilized ResNet50 for crack detection in underwater concrete structures, proving its robustness in complex scenarios such as low visibility and image distortions. Moreover, [14] implemented a Vision Transformer model for crack detection and achieved higher accuracy than ResNet50 on a diverse dataset of concrete surfaces with cracks, highlighting the strength of ViTs in handling wide variations in crack morphology [22]. These findings underscore the potential of integrating transformer-based models with CNNs for enhanced detection performance and adaptability in real-world conditions. Future research is expected to focus on combining these models with real-time inspection systems and UAV-based platforms for automated structural monitoring [20].

3. Methodology

The proposed framework treats crack detection as a multimedia image analysis task, where images are processed, transformed, and classified using deep learning architectures. By framing crack detection as a multimedia problem, the approach leverages both spatial and contextual information in images for improved accuracy. The pipeline includes dataset preparation, multimedia content preprocessing, model

training, and performance evaluation. Each stage of the pipeline is designed to ensure the integrity of data flow and robustness of the predictive models.

3.1 Materials

The tools used in this study are Google Colab Notebook and a couple of Python programming language libraries. Table 1 shows the technologies used and their uses. The cloud-based Colab environment provided scalable GPU resources, enabling efficient model training on large datasets.

Table 1. *ML steps and the technologies used*

Machine Learning Step	Technology used
Data collection	Kaggle, Shutterstock, and iStock
Data processing	Numpy,
Data Visualization	Matplotlib, Seaborn, Plotly. Express
Data pre-processing	Pathlib, Sklearn.model_selection,
Model training	ViT and Resnet50, Pytorch, tensorflow
Model evaluation	Adam optimizer

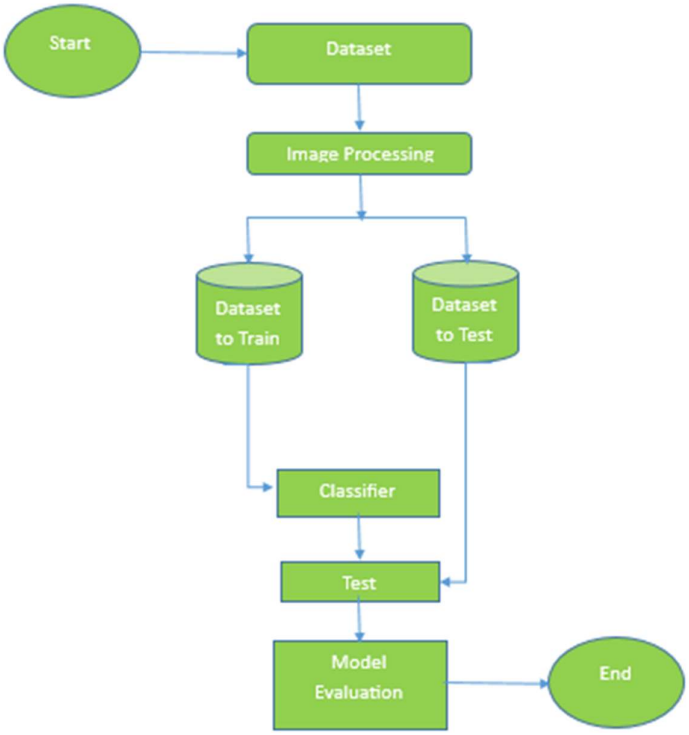

Fig. 1. *Flow-chart of a model*

3.2 Architecture of the Proposed Models

The models used were ViT and ResNet50 for the detection of concrete cracks. Each model had 5 stages. Initially, the images (i.e. concrete structures) were passed into the model at the first stage. This step ensured that the raw image data could be efficiently loaded and preprocessed for downstream analysis. In the second stage, the data frame containing the file path was generated. Generating a structured data frame helped maintain proper indexing and metadata tracking for each image, which is essential for training and evaluation. In the third stage, the images were classified and segmented into two hairline cracks and structural cracks in a separate folder for image classification. Segmentation ensured that the model could focus on the relevant areas of cracks, improving feature extraction and learning. The fourth stage included model training, which tried to familiarize the model with concrete crack images. During training, data augmentation techniques were applied to increase dataset variability and reduce overfitting. Model validation aimed to understand how efficient the model was, and model testing aimed to understand how effective the models were in processing and identifying types of cracks. The fifth stage was model evaluation, in this stage, the two models were compared to check their performance in concrete crack

detection. Performance metrics such as accuracy, precision, recall, F1-score, and confusion matrices were used for comprehensive model evaluation.

3.2.1 ViT Architecture

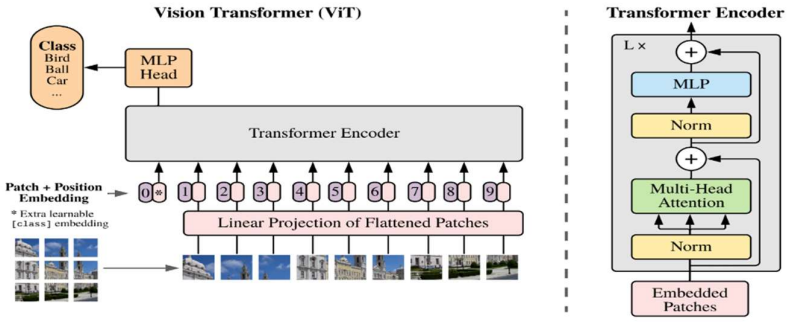

Fig. 2. *Architecture of the vision transformer*

Vision Transformer treats images as sequences of patches and applies self-attention mechanisms to capture long-range dependencies between image patches, processing each patch as a token by allowing the model to capture both local features such as small cracks and global contextual information like concrete texture. The self-attention mechanism enables ViT to weigh different regions of the image according to their relevance, which improves detection of subtle cracks in complex surfaces.

3.2.2 ResNet50 Architecture

ResNet-50 is a deep convolutional neural network with 50 layers, built using residual blocks that introduce skip connections, allowing the network to learn identity mappings and combat the vanishing gradient problem. It uses a bottleneck architecture where each residual block consists of three layers: 1x1, 3x3, and 1x1 convolutions. The network is structured into five stages, each with multiple residual blocks, followed by global average pooling and a fully connected layer for classification. The input image size is typically 224x224, and the network has around 23 million parameters. ResNet50's deep architecture allows it to capture fine-grained features in crack patterns while maintaining stability during training.

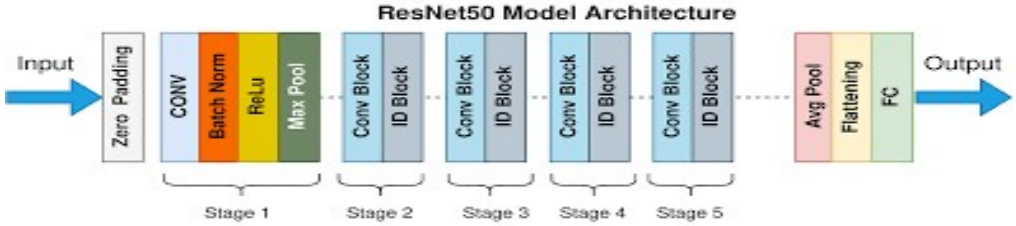

Fig. 3. *Architecture of the ResNet50*

3.3 Data Collection

Image data collection is an important aspect of developing an effective crack detection machine learning model. The quality and quantity of the collected images significantly impact the models' performance. Previous research studies have noted that the quantity and quality of images determine how fast and effective a model can predict cracks. Careful annotation of images with metadata such as crack type, length, and location ensures accurate supervised learning.

In this study, the collected data were primarily in the form of images. Each image was annotated with relevant metadata, such as the type of crack and the dimensions of the crack. A total of 10,000 crack images were collected and used, which consisted of 5,000 Hairline cracks and 5,000 Structural cracks images. This balanced dataset allows the models to learn equally from both classes, reducing bias in predictions.

3.4 Data Pre-Processing

Data pre-processing is required in developing an accurate machine-learning model for predicting concrete cracks. It is an important step in the data analysis process which involves data cleaning, transformation, and organizing raw data to make it suitable for further analysis or modeling. Preprocessing steps will

include resizing images, normalizing pixel values, Image Classification and Segmentation, and augmenting data to increase the dataset size and variability. Augmentation techniques included rotation, flipping, brightness adjustment, and noise addition to simulate real-world inspection scenarios. This process ensured the accuracy, consistency, and relevance of data, thereby improving the quality and reliability of the predictions derived from it.

3.5 Data Visualization

Data visualization is usually carried out to determine whether data is properly loaded from its directories. During visualization, defects such as blurry images, irrelevant objects, or misclassified data that could negatively affect model performance are observed. Visualization also helps in understanding the distribution of crack types, image resolutions, and potential biases in the dataset.

3.6 Model Training, Validation and Testing

Models are trained and tested using algorithms. An ML algorithm is a mathematical model or set of rules used by a computer to learn patterns from data and make predictions or decisions without being explicitly programmed for specific tasks. ML algorithms can analyze complex data patterns and make accurate predictions. In the context of this study, ResNet50 and Vision Transformer machine learning algorithms are explored. Training involved splitting the dataset into training, validation, and test sets to ensure unbiased evaluation of model performance.

3.6.1 ViT

ViT is a deep learning model that applies the Transformer architecture, originally designed for natural language processing, to image data for tasks like image classification. Instead of relying on convolutional layers like CNNs, ViT processes images by dividing them into smaller patches and treating them as tokens (similar to words in NLP). This enables ViT to capture both local and global dependencies, which is critical for detecting cracks that vary in size, orientation, and texture. This approach enables the model to capture long-range dependencies and global relationships within an image more effectively than CNNs. Additionally, ViT can be combined with transfer learning from pre-trained models to improve performance on limited datasets.

Mathematical Algorithm:

1. Image Patch Extraction:

$X \in R^{H\times W\times C}$ (input image)

Split into $N = \frac{H\times W}{P^2}$, patches of size P x P x C, where H, W, and C are the height, width, and channels, respectively.

2. Linear Projection of Patches:

$$z_o = [x_{cls}; x_1\mathbf{E}; x_2\mathbf{E}; \dots; x_N\mathbf{E}] + \mathbf{E}_{pos}$$

Where $x_i \in R^{P^2.C}$ are flattened patches, $\mathbf{E}$ is the learnable embedding matrix, and $\mathbf{E}_{pos}$, is the positional encoding.

3. Self-Attention (Multi-Head):

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where $Q = zW_Q, K = zW_K, V = zW_V$ are queries, keys, and values derived from the input embeddings z, and 4 is the dimension of the key vectors.

4. Feedforward Network:

$$FFN(z) = MLP(LayerNorm(z)) + z$$

5. Output Classification:

$$y = MLP(CLS - token\ output)$$

where the output of the classification token is used for the final prediction.

3.6.2 ResNet-50

ResNet-50 is a deep CNN architecture designed to tackle the degradation problem in deep networks by introducing **residual connections**. These residual shortcuts effectively address vanishing gradients, ensuring stable training even as the network depth increases. These connections allow the network to learn identity mappings, making it easier to train very deep networks. By reusing features from earlier layers, the model enhances its ability to detect both fine-grained cracks and large structural discontinuities. ResNet-50 consists of 50 layers, organized into residual blocks, where the input to a block is added directly to its output, allowing the model to maintain information and improve gradient flow. This

architecture has been widely adopted in computer vision tasks and has demonstrated superior performance in structural defect detection compared to traditional CNNs.

Mathematical Algorithm:

1. **Residual Block**:

For a given input x, the residual block computes:

$$\hat{x} = F(x, \{W_i\}) + x$$

where $F(x, \{W_i\})$ is the function representing the stacked convolutional layers and TV; are the weights. The input x is added directly to the output $F(x, \{W_i\})$, forming the residual connection.

2. **Bottleneck Residual Block:**

$$F(x) = W_3 \cdot \sigma(W_2 \cdot \sigma(W_1 \cdot x))$$

where:

- W_1 is a lx1 convolution for reducing dimensionality.
- W_2 is a 3x3 convolution for feature extraction.
- W_3 is a lx1 convolution to restore dimensionality.
- σ is the activation function (ReLU).

3. **Final Classification:**

After passing through the network, the feature map is reduced using global average pooling and fed into a fully connected layer for final classification:

$$y = softmax(Wx)$$

where W are the weights of the fully connected layer. The softmax activation ensures that the output is normalized into probability distributions across the target classes, making it suitable for multi-class crack classification.

ResNet-50's use of residual connections makes it highly effective at training deep networks while avoiding the vanishing gradient problem, enabling more accurate image recognition tasks.

3.7 Model Evaluation

Adam's evaluation metric was employed to assess the accuracy and performance of the adopted machine-learning models for predicting concrete cracks. The choice of Adam was motivated by its adaptive learning rate mechanism, which adjusts step sizes for each parameter individually and improves convergence speed. AOA is mostly used by researchers in machine-learning-based projects to identify and predict models, especially for optimized outcomes or for evaluation of the existing models for better outcomes. It is mostly adopted for its rapidness and robustness. It requires fewer processes than other algorithms and is thus preferred by researchers.

Table 3.2 The pseudo-code for the adopted algorithm

Adam Optimizer Algorithm: Input
Step 1: Initialize the process, $a_b = 0$ and x_1
Step 2: For, n = 1, …N, do
$a_n = S_{1,n} a_{n-1} + (1 - \rho_{1,n}) g_n$
$\widehat{\nabla}_N = h_N(g_1, g_2, \ldots, g_n)$
$x_{n+1} = x_n - \frac{m_n b_n}{\sqrt{\widehat{\nabla} n}}$
Model evaluation

4. Result and Discussion

The developed model was experimented by applying the Pre-trained Vision Transformer model on the datasets. The model was trained and tested and the outcomes obtained are stored. The choice of using pre-trained models significantly reduced computational cost and accelerated convergence, as the models could leverage prior knowledge from large-scale datasets. Also, the experimentation was carried out on the Pre-trained ResNet50 model and the outcomes are stored for comparison against the ViT model. This comparative analysis helps in understanding how transformer-based models differ from traditional CNNs in terms of feature extraction and robustness against noise in concrete crack images. Through experimentation, the outcomes from the models were evaluated and compared for better-performing models with higher accuracy. The evaluation included not only accuracy but also loss convergence, precision, recall, and F1-score, ensuring a comprehensive assessment of model performance.

4.1 Loss Function (graph) of the Vision Transformer Model

Fig. 4 illustrates the loss curves for training and validation, sets over a series of epochs. The **training loss** decreases smoothly, indicating that the model learns and fits well with the training data. This steady decline in training loss reflects the ViT's ability to capture fine-grained details in crack patterns by leveraging its attention mechanism. However, the **validation loss** displays significant fluctuations and remains consistently higher than the training loss. This suggests that while the model performs well on the train data, it may be overfitting, as it struggles to generalize effectively to the validation sets. Such overfitting is common in deep models when the dataset is not sufficiently diverse, highlighting the need for data augmentation or regularization techniques. The instability in the validation loss further indicates potential difficulties in making reliable predictions on unseen data.

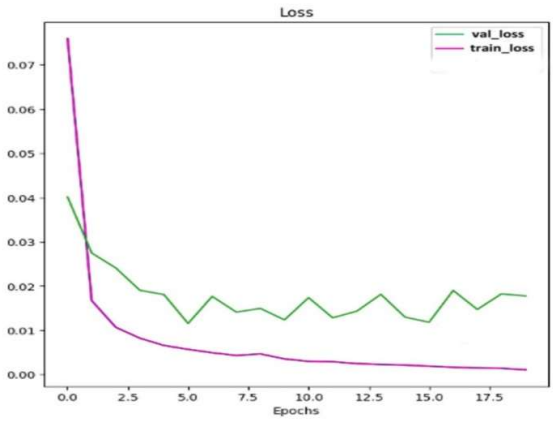

Fig. 4. *ViT loss function graph*

4.2 Loss Function (graph) of the ResNet50 Model

The validation loss decreases smoothly and approaches near zero in Fig. 4, showing strong performance on the training data. This trend reflects the effectiveness of convolutional feature extraction in ResNet50, which is particularly well-suited for capturing spatial hierarchies in crack patterns. Although the validation loss remains higher in Fig. 5 than the training loss, the overall gap is smaller, which is a good sign that the model generalizes better to unseen data compared to Fig. 4. This indicates that ResNet50 is less prone to overfitting compared to the Vision Transformer, making it more stable for real-world applications where unseen data variability is common.

Fig. 5. *ResNet50 loss function graph*

The validation loss decreases smoothly and approaches near zero in Fig. 4, showing strong performance on the training data. Such performance suggests that ResNet50 can effectively balance training accuracy with validation reliability, demonstrating its robustness in handling complex datasets of concrete cracks. Although the validation loss remains higher in Fig. 5 than the training loss, the overall gap is smaller, which is a good sign that the model generalizes better to unseen data compared to Fig. 4. This makes ResNet50 a promising candidate for deployment in structural health monitoring systems, where both precision and generalization are crucial for decision-making.

4.3 Model Accuracy

The Accuracy of the model is calculated based on the given formula;

$$\textbf{Accuracy} = \frac{True\ Positive + True\ Nevative}{True\ Positive + False\ Positive + True\ Negative + False\ Negative}$$

This metric provides a straightforward measure of the model's performance by quantifying the proportion of correct predictions out of the total cases evaluated.

4.4 Accuracy (graph) of the Vision Transformer Model

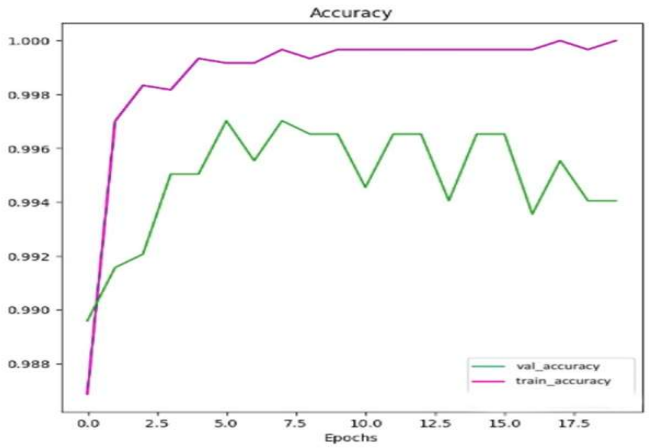

Fig. 6. *Accuracy graph of ViT model*

In **Fig. 6,** both the **training accuracy** and **validation accuracy** approach near-perfect values (close to 100%), but the patterns are distinct. This suggests that the Vision Transformer has strong representational capacity, enabling it to quickly adapt to the training dataset. The **training accuracy** rapidly increases and remains stable across epochs, indicating that the model fits the training data well. Such stability in training accuracy reflects the ability of ViT to capture complex crack features effectively. Although, the **validation accuracy** fluctuates noticeably after the initial rise. These fluctuations highlight potential overfitting issues, as the model may be memorizing training patterns rather than generalizing robustly to unseen examples. While the model learns the training data very well, its ability to generalize to unseen data is less consistent due to the fluctuations in the validation accuracy.

4.5 Accuracy (graph) of the ResNet50 Model

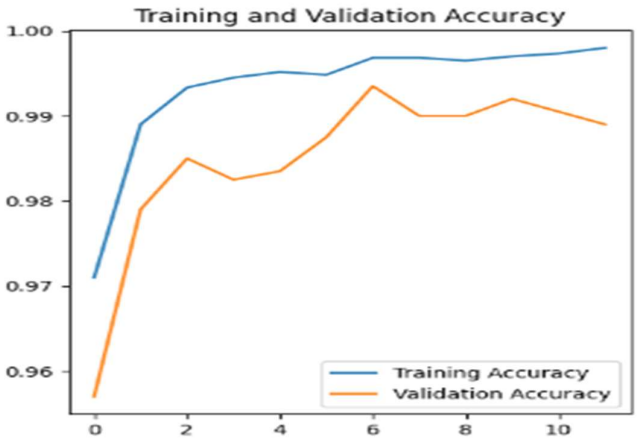

Fig. 7. *Accuracy graph of ResNet50 model*

From Fig. 7, the **training accuracy** increases rapidly, reaching nearly 100% within the first few epochs, and then stabilizes. This indicates that ResNet50 is able to efficiently learn relevant crack features from the dataset with fewer fluctuations compared to ViT. The **validation accuracy** follows a similar trend, starting lower but gradually improving and maintaining a high value with minimal fluctuation. Such consistency between training and validation accuracy curves demonstrates the robustness of ResNet50 in handling both training and unseen validation data. The close alignment between the training

and validation accuracy curves suggests that the model is performing well, with minimal overfitting and good generalization to the validation data. This result highlights that ResNet50 may be more suitable than ViT in real-world scenarios where stable generalization across diverse crack images is critical.

4.6 Experimental Results of the ResNet50 and the ViT model

The prediction was made for the ResNet50 and the ViT models with the concrete cracks data and the following outcomes were obtained: the validation and training results are shown in Table 2 and Table 3. These tables provide a comparative overview, enabling clear benchmarking of the two models across identical experimental conditions.

Table 2. *The experimental Result of ResNet50*

Epoch	Training Loss	Training Accuracy	Validation Accuracy	Validation Loss
1	0.1741	0.9252	0.9570	0.1057
2	0.0321	0.9876	0.9790	0.0476
3	0.0279	0.9926	0.9850	0.0327
4	0.0109	0.9954	0.9825	0.0400
5	0.0138	0.9949	0.9835	0.0353
6	0.0140	0.9954	0.9875	0.0273
7	0.0091	0.9970	0.9935	0.0141
8	0.0090	0.9966	0.9900	0.0219
9	0.0107	0.9965	0.9900	0.0164
10	0.0081	0.9976	0.9920	0.0157
11	0.0074	0.9977	0.9905	0.0224
12	0.0101	0.9973	0.9890	0.0189

Table 3. *The experimental Result of the ViT model*

Epoch	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy
1	0.0869	0.9814	0.0946	0.9563
2	0.0180	0.9970	0.0985	0.9529
3	0.0115	0.9982	0.1127	0.9504
4	0.0088	0.9982	0.1173	0.9489
5	0.0070	0.9993	0.1224	0.9489
6	0.0060	0.9992	0.0942	0.9563
7	0.0052	0.9992	0.1302	0.9454
8	0.0046	0.9995	0.1190	0.9499
9	0.0049	0.9993	0.1278	0.9479
10	0.0038	0.9997	0.1178	0.9514
11	0.0032	0.9997	0.1450	0.9425
12	0.0031	1.0000	0.1229	0.9504
13	0.0027	0.9997	0.1338	0.9469
14	0.0025	1.0000	0.1620	0.9405
15	0.0023	0.9997	0.1265	0.9474
16	0.0020	0.9997	0.1207	0.9504
17	0.0018	0.9997	0.1607	0.9410
18	0.0016	1.0000	0.1384	0.9459
19	0.0015	0.9997	0.1622	0.9400
20	0.0012	1.0000	0.1521	0.9425

Table 3 shows the outcome of the training and validation data of the Vision Transformer model when it is trained with a data set of 10,000 images which consist of 5,000 hairline cracks and 5,000 structural cracks with 20 epochs given to the model at a learning rate of lr=5e-4. The balanced distribution of hairline and structural crack images ensured that both fine and major crack types were adequately represented in the dataset, supporting a fair evaluation. The ResNet50 model was also trained on the same data sets with a learning rate of lr=0.001. This adjustment highlights the sensitivity of ResNet50 to learning rate configurations, as an inappropriate choice can lead to unstable convergence or suboptimal accuracy. The reduction in the ResNet50 learning rate is due to its low performance when trained with a higher learning rate of 5e-4 with 20 epochs.

4.7. Comparative Analysis of the Models Developed

The performances of both models are evaluated through metric evaluation techniques. The outcomes are (refer to Table 4 and 5): These metrics include training accuracy, validation accuracy, training loss, and validation loss, which together provide a holistic understanding of model efficiency and generalization.

Table 4. *Comparative analysis of the models for Accuracy*

Model	Vision Transformer	ResNet50
Train Accuracy	99.97%	99.93%
Test Accuracy	99.40%	99.80%

Table 5. *Comparative analysis of the models for the Loss function*

Model	Vision Transformer	ResNet50
Train Loss	0.0012	0.0016
Test Loss	0.0178	0.0011

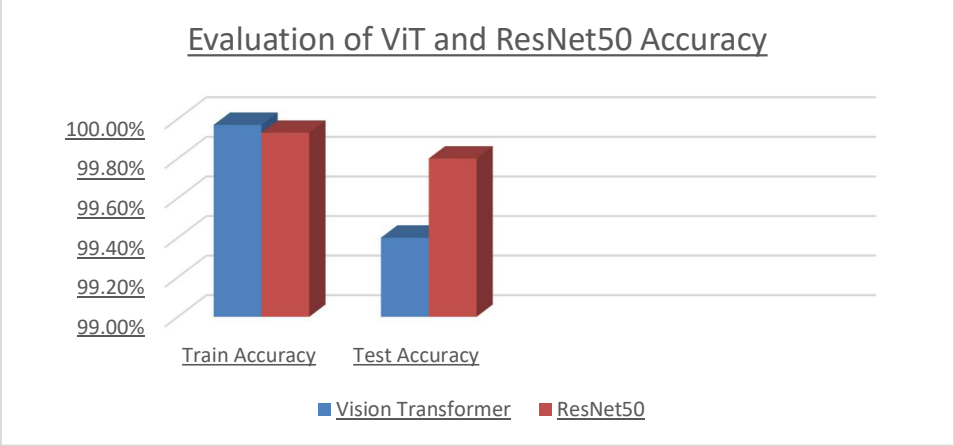

Fig. 8. *Comparative analysis of ViT & ResNet50 accuracy*

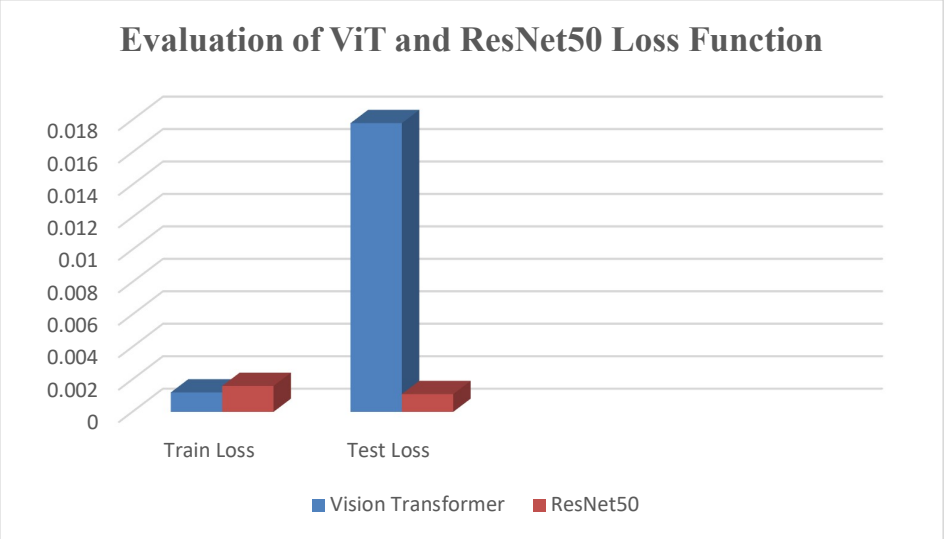

Fig. 9. *Comparative analysis of ViT & ResNet50 loss function*

Fig. 8 illustrates the comparative results of the accuracy of the ViT (vision transformer) and ResNet50 model, while Fig. 9 shows their loss function. Visualizing the metrics side by side makes it easier to highlight the strengths and weaknesses of each model across different evaluation criteria. From the observation, it is understood that the outcomes obtained from the ResNet50 model for identifying and detecting the hairline concrete cracks and structural concrete cracks are better than the ViT model. This superiority of ResNet50 can be attributed to its residual connections, which enhance feature propagation and improve stability during deeper training.

The evaluation metrics show that train accuracy is higher, but the test accuracy is lower in the ViT compared to the ResNet50 model. This gap between training and test performance suggests that ViT, despite its strong representational power, may be more prone to overfitting when applied to relatively smaller datasets such as the one used in this study. Also, the training loss is lower in ViT compared to the ResNet50 which gives it a good chance of prediction, although the test loss in ViT is much higher compared to the ResNet50 which indicates how well the ResNet50 model is, in detecting hairline or structural cracks in a concrete structure. This finding confirms that while ViT can learn effectively within training

boundaries, ResNet50 demonstrates stronger generalization in real-world conditions, making it a more reliable choice for automated crack detection systems integrated into UAVs or robotic inspections.

5 Discussion

The results show that while both ResNet50 and ViT achieved high accuracy in detecting hairline and structural cracks, ResNet50 demonstrated better generalization and stability, supporting the hypothesis that CNN-based models are more robust on medium-sized datasets. This aligns with prior studies that emphasize the reliability of residual networks in noisy or complex environments, while also confirming that ViTs, though powerful, require larger datasets and careful tuning. An unexpected finding was ViT's near-perfect training accuracy despite fluctuating validation accuracy, likely due to overfitting and limited dataset diversity. Practically, this suggests ResNet50 is more suitable for immediate industry deployment in UAV or robotic inspections, whereas ViTs hold future promise for academia if supported by larger datasets and hybrid approaches. Nonetheless, the study is limited by its focus on only two models, reliance on image-based data, and exclusion of other advanced architectures, which should be addressed in future research.

6 Conclusion and Recommendations

6.1 Conclusion

The current study aimed at applying the ResNet50 model and ViT model on datasets of concrete cracks images. This study provides insights into the comparative effectiveness of convolutional versus transformer-based architectures in detecting structural defects. By leveraging deep learning techniques, the research contributes to the ongoing efforts to automate structural health monitoring, which is critical for ensuring the safety and longevity of infrastructure. The research examined the performance metrics, highlighting that both models achieved exceptionally high accuracy, indicating their suitability for automated crack detection in civil infrastructure. The close performance between the two models suggests that either approach can be reliably deployed, although specific deployment scenarios may favor one model over the other based on computational resources or training dataset characteristics. The ViT model acquired a higher training accuracy of (99.97%) than the ResNet50 model, which gave (99.93%), demonstrating the ViT model's slightly superior learning capability during training. This indicates that transformer-based models may have an advantage in capturing complex patterns in image data, potentially due to their attention mechanisms that allow them to focus on key image regions. Although their respective testing accuracy was (99.40%) for ViT and (99.80%) for the ResNet50 model, these results suggest that while ViT excels in training, ResNet50 generalizes slightly better on unseen data, emphasizing the importance of balancing training and testing performance in model selection. Such findings underline the practical significance of considering both overfitting tendencies and model robustness when implementing AI solutions in real-world structural monitoring tasks. Future research could explore hybrid approaches, combining the strengths of convolutional and transformer-based models, or investigate the effect of larger and more diverse datasets to further improve detection performance.

6.2 Limitations

The study only compared the ResNet50 model against the ViT model and thus other architectures are neglected. This limited scope may reduce the generalizability of the findings across different deep learning frameworks. Also, little augmentation was made on the ResNet50 to make it overcome the challenge of overfitting during the experimentation which was not done on the ViT model.

6.3 Recommendations

The current study developed a ResNet50 model and compared the results with the ViT model. Future studies should explore a broader range of architectures to provide a more comprehensive understanding of model effectiveness in crack detection. However, in the future, the same could be extended where instead of the ResNet50 model other models like ResNet, LeNet, AlexNet, and other architectures without CNN layers could be used for comparing the outcomes with the ViT or ResNet50 model. Similarly, future research could focus on models of clustered crack prediction in concrete structures by examining their precision, accuracy, recall, and f1-score to find the best model for predicting and detecting structural crack types. This could lead to the development of more reliable predictive maintenance systems for infrastructure. In the future, more CNN-based architectures can be analyzed and studied for performance

and accuracy. Different evaluation metrics could also be utilized to compare the performances of the models and to find the better model and metric.

Furthermore, different NN like RNN (Recurrent NN), and ANN (Artificial NN) could also be focused on finding better NN models in crack detection, than ResNet50 and vision transformers. Exploring temporal or sequential models like RNNs may be particularly useful if monitoring changes in crack patterns over time is required.

6.4 Contribution to Knowledge

This study contributes to knowledge by its ability to effectively detect hairline cracks compared to other methods such as visual inspection, measurements of crack width, on-site stress examinations, ultra-sonic, image processing techniques, *etc.* The adoption of deep learning models provides a faster and more scalable alternative, which is especially valuable for large-scale infrastructure monitoring Its high accuracy in detecting and classifying cracks in structures will reduce labor costs during inspection, which makes it cost-effective.

Furthermore, this technique of crack detection can easily be integrated into a drone or robot, which will also reduce risk during crack inspection in building structures.

7. Acknowledgement

I extend sincere appreciation to the Head of the Civil Engineering Department, Prof. O.J. Oyedepo, and my supervisor, Dr. O.O. Omotayo. Your roles have gone beyond mentorship you have been sources of constant care, guidance, and unwavering support. Your belief in our potential has propelled us to reach greater heights, and I am truly thankful for the invaluable lessons and inspiration you have provided. I will also extent my appreciation to my friend, David John for his immense support for success of this research.

Compliance with Ethical Standards

Conflicts of interest: Authors declared that they have no conflict of interest.

Human participants: The conducted research follows the ethical standards and the authors ensured that they have not conducted any studies with human participants or animals.

Declaration of generative AI and AI-assisted technologies in the writing process: The authors declare that no generative AI or AI-assisted technologies were used in the writing process of this manuscript.

References

[1] Ali, R., Chuah, J. H., Talip, M. S. A., Mokhtar, N., & Shoaib, M. A. (2022). Structural crack Detection using deep convolutional neural networks. *Automation in Construction*, 133, 103989.https://doi.org/10.26682/sjuod.2017.20.1.67

[2] Askar, M. K., AL-Kamaki, Y. A. M. A. N., Ferhidi, R., & Majeed, H. K. (2023). Cracks in Concrete structures causes and treatments: A Review. *Journal of Duhok University,* 26(2), 148-165. https://doi.org/10.5897/ijps11.1181

[3] Shull, P. J. (2002). *Nondestructive evaluation: Theory, techniques, and applications*. Marcel Dekker.

[4] Montalvão, D., Maia, N. M. M., & Ribeiro, A. M. R. (2006). A review of vibration-based structural health monitoring with special emphasis on composite materials. *Shock and Vibration Digest, 38*(4), 295–326. https://doi.org/10.1177/0583102406066290

[5] Nadernejad, E., Sharifzadeh, S., & Hassanpour, H. (2008). Edge detection techniques: Evaluations and comparison. *Applied Mathematical Sciences, 2*(31), 1507–1520

[6] Batta, M. (2019). Machine learning algorithms – A review. *International Journal of Science and Research*, *9*(1), 381–386. https://doi.org/10.21275/ART20203995

[7] Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

[8] Cha, Y. J., Choi, W., & Büyüköztürk, O. (2017). Deep learning-based crack damage detectionUsing convolutional neural networks. *Computer-Aided Civil and Infrastructure Engineering*. https://doi.org/10.1016/j.clay.2008.11.007

[9] Zhang, L., Yang, F., Zhang, Y. D., & Zhu, Y. J. (2016). Road crack detection using deep Convolutional neural network. *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. https://doi.org/10.1088/1757-899x/1197/1/012082

[10] Li, Y., Liu, Y., Liu, Y., & Xia, Y. (2018). Transfer learning-based automatic detection of Diabetic retinopathy in retinal fundus photographs. *Proceedings of the 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. https://doi.org/10.1007/s12205-013-1241-9

[11] Zhang, J., Zhang, J., & Li, Q. (2019). "Concrete crack detection based on deep learning model." *Journal of Structural Engineering*, 145(5), 04019036.
[12] Hamada, K., et al. (2020). "Underwater concrete crack detection using deep residual learning."*Sensors*, 20(18), 5300.
[13] Chen, S., Zhang, H., & Liu, Y. (2022). "A Vision Transformer approach for crack detection in concrete structures." *Automation in Construction*, 135, 104151.
[14] Chun, P. J., Izumi, S., & Yamane, T. (2021). Automatic detection method of cracks from Concrete surface imagery using a two-step light gradient boosting machine. *Computer-Aided Civil and Infrastructure Engineering,* 36(1), 61-72.
[15] Flah, M., Suleiman, A. R., & Nehdi, M. L. (2020). Classification and quantification of cracks In concrete structures using deep learning image-based techniques. *Cement and Concrete Composites.*
[16] Golewski, G. L. (2023). The phenomenon of cracking in cement concretes and reinforced concrete structures: *the mechanism of cracks formation, causes of their initiation, types and Buildings* . https://doi.org/10.1016/j.matpr.2022.04.210
[17] Hassani, S. & Dackermann, U. (2023). A systematic review of advanced sensor technologies For non-destructive testing and structural health monitoring. *Sensors.*
[18] Kheradmandi, N. & Mehranfar, V. (2022). A critical review and comparative study on image Segmentation-based techniques for pavement crack detection. *Construction and Building Materials. https://doi.org/10.1061/(ASCE)mt.1943-5533.0000782*
[19] Liu, X., Gao, Y., Ma, L., Zhang, Z., Zhang, S., & Nie, W. (2019). Generative adversarial networks For synthetic data generation in machine fault diagnosis. *IEEE Access.*
[20] Yuan, J., Mao, W., Hu, C., Zheng, J., Zheng, D., & Yang, Y. (2023). Leak detection and Localization techniques in oil and gas pipeline: *A bibliometric and systematic review. Engineering Failure Analysis,* 146, 107060.
[21] Zawad, M. R. S., Zawad, M. F. S., Rahman, M. A., and Priyom, S. N. (2021). A comparative review Of image processing-based crack detection techniques on civil engineering structures. *Journal of Soft Computing in Civil Engineering*, 5(3), 58-74.
[22] Li, Y., Zhao, X., & Peng, L. (2023). "Comparison of Vision Transformer and CNN models for crack detection in satellite imagery of bridges." *IEEE Transactions on Geoscience and Remote Sensing*, 61, 5103521.