REVISTA MATÉRIA V.28 N.1
ISSN 1517-7076 articles e20230010, 2023

Application of machine learning for crack detection on concrete structures using CNN architecture

P. Padmapoorani[1], S. Senthilkumar[1]

[1]KSR College of Engineering, Department of Civil Engineering, Namakkal, Tamilnadu , 637215, India
e-mail: padmapooranip@ksrce.ac.in; senthil.env@gmail.com

ABSTRACT

Cracks in concrete structures are caused due to contraction and expansion irregularities, from potential damages caused in the buildings. These irregularities and damages are assessed by the engineers manually or through identification and prediction models with machine learning techniques to evaluate the impact and significance of the structural health in buildings. This research aims at applying machine learning based on VGG16-Net model for the detection of cracks in concrete structures. The proposed model is of CNN (convolutional neural network) + VGG neural network based architecture. The study uses the gradient boosting algorithm for image segmentation. The datasets are obtained from "Kaggle" resource and the library used is 'Hugging Face Transformers'. To evaluate the developed models' performance metrics such as "accuracy, precision, recall and f1-score" are used. The 'accuracy' score obtained is compared against the 'ViT' (Google transformer) accuracy rate, for comparison. The proposed model achieved 98% validation accuracy rate with 0.3% loss. Thus the developed research contributes an innovative and a novel ML model that predicts and identifies the cracks in concrete structures with less loss and higher accuracy with CNN architecture than ViT (vision transformer) models. Current study also provides more input upon CNN being more accurate than ViT models for future researchers for comparative analyses.

Keywords: VGG16-Net; concrete cracks; concrete non-cracks; convolutional neural network; structural health; concrete structures; ViT; vision transformers.

1. INTRODUCTION

In the recent years the researches on the structural health of concrete has increased rapidly. The study of infrastructures on the dams, buildings, bridges and roads have been done by the investigators for the health, impact, mishandlings and other structural observations [1]. The expansion and contraction of the concrete structures in the buildings have been observed as the primary reason to study the health of the concrete structures. To identify these damages, cracks and breaks in the concrete structures the investigators have been adopting and leaning towards machine learning models in the recent years than observing through direct and physical investigation that requires more time, budget, labour and human intervention.

The machine learning based models in the concrete structural health analysis has been focused by researchers to study the variations in the expansions, vibrations, features (frequency and spatial), contractions and dampness [2]. Automatic detection of the concrete structures' cracks and how it impacts the health of the concrete structures have been recently focused with the "vision transformers" (ViT) as the core focus.

The studies by authors [3] and [4] focused on pavement cracks through ViT models where they insisted that the Convolutional Neural-Networks (CNNs) as the base model network for identifying the cracks in pavements which is evaluated and presumed to be the better and higher performing models in image segmenting and classifying.

The CNN models in identifying the cracks in the images of pavements and concrete structures have been identified as better model. However, study by [5] insisted that deep learning-based CNN models in detecting the cracks with the aerial support of unmanned vehicles (UAV) as a combination as dual process is faster, robust and reliable than detecting cracks in the images of random inputs. Digital images in the automated crack detection models as inputs than videos in real-time have been attempted more to evaluate the detection models. The measurement of the cracks in the concrete structures are calculated by considering various features, like width, length, thickness, thinness, mass and depth of the crack [6]. Though the manual crack detection is measurable

Corresponding Author: P. Padmapoorani Received on 13/01/2023 Accepted on 27/02/2023

DOI: https://doi.org/10.1590/1517-7076-RMAT-2023-0010

it is not always a reliable outcome since there is a huge possibility of error and miscalculations through human interventions. Henceforth, automatic detection of cracks in concrete structures and pavements through deep learning, where advanced techniques like transformers are recently identified as 'highly' performing models than the CNN models with better accuracy [7]. The models traditionally adopted NN models are ResNet50 [8], AlexNet [9], VGG16-Net [10], LeNet [11], GoogleNet [12,13], MobileNets [14] and DenseNet models [15].

Initially vision transformers (ViT) were applied only in the natural language processing (NLP) by replacing the RNN (recurrent neural-network) models with long-short-term memory (LSTM) approach. Later they were utilized in question answering, language translations and text classifications [16] applications too. ViT models have been tremendously dominating in SOTA (State-Of-The-Art) performance and efficiency with the NLP datasets, than other categories. Similarly the vision transformers are also popularly used for the advantages over CNN like its cost, processing time and speed. However there exists a debate that in certain areas, CNN models are more efficient with accuracy than the transformers. The ViT thus could be assumed as, a major contributor in the computer vision domains that is adopted for different applications like image segmentation, video understanding, object detection and image detection in machine learning.

In this research the author will develop a traditional CNN model to detect cracks in the concrete structures to examine its health with the main purpose as: "*applying machine learning on concrete structures to monitor and examine the health of concretes*". Simultaneously the research will also analyse and examine the objectives:

- To examine the accuracy achieved by a conventional model in detecting the cracks in concrete structures;
- To examine the accuracy achieved by a contemporary model in detecting the cracks in concrete structures;
- To compare the outcomes to weigh the most appropriate model based on cost, time, resource, accuracy rate through metric evaluation.

2. LITERATURE REVIEW

The studies on the convolutional networks in detecting cracks in concrete structures and monitoring the concrete structural health are primarily focused. Models that use CNN architecture in deep learning and vision based (i.e. digital image segmentation and vision transformers) are studied as secondary studies as literary resources.

2.1. Structural health monitoring and crack detection in convolutional networks

Authors [17, 18], conducted an inspection on concrete structures to study how the cracks appear through CNN architecture. Authors [17] developed a deep CNN model that achieved 92% f1-score and recall. Similarly, the model developed by the authors [18] achieved 99% accuracy with image processing technique and machine learning algorithm. The developed model was based on ANN (artificial neural-network). The samples used were 1000 images. Authors [19] used multi-resolution analysis (ResNet50 and AlexNet) as their data analysis technique where the deep learning is used. Their model achieved 90% accuracy with CNN architecture with 56000 images. The outcomes were compared against SDNET2018 model developed by [20]. From all these models and the samples size it could be observed that, the models majorly used the CNN architecture for better accuracy and the size of the sample datasets varied from 1000–56000 images. However, the larger the data the longer the processing of dataset in computing and the cost incurs. Hence, it is advised by the authors in their studies that, sample should be lesser for more accurate and robust crack detection in concrete structures.

By adopting a hybrid machine learning approach in their study authors [21] developed a model with SVM (support vector-machine) and CNN architecture that uses the aerial based unmanned vehicle (UAV). The model achieved 92% accuracy than the single classifier and image processing methods in crack detection to monitor concrete structural health. Authors [22] used PCA (principal component analysis) with four machine learning algorithms (gradient booster, decision tree, AdaBoost and randomized tree) in their model. Though they found that AdaBoost as effective algorithm. The gradient booster and randomized tree algorithms were recorded with overfitting issues that ought to be prevented in future studies for higher accuracy. From these studies it can be deduced that, adopting the decision and random tree algorithms will cause overfitting issue, thus prior analysing the data researcher should consider about adopting the better algorithm for predicting the cracks.

2.2. Traditional methods versus contemporary methods

Several existing researchers [23–28] have studied about the traditional methods in the crack detection in buildings and pavements. According to the common findings of the studies, it has been claimed that the traditional methods are better in accuracy than the transformer models. The traditional methods in crack detection uses ResNet, AlexNet, MobileNet, Inception, Convolutional, DenseNet and LeNet models that has more layers than the transformers. However, the accuracy of the outcomes was predicted to be 99% where the error rate is at 1%.

Contrarily, several authors [1–3,29,30] focused and examined exclusively upon the vision based crack detection models. They claimed that though the traditional models are better in accuracy than the vision transformer models that averagely produce outcomes that are of 95% accurate, the transformer models are rapid, robust, incurs lesser costs and lesser time for computing and processing. Hence, the recent researchers have been adopting the vision transformers as crack detection models that produce lesser accuracy than traditional models in micro cracks, macro cracks, complex and closely-spaced crack detection in buildings, pavements and asphalt blocks that are of concrete structures. To monitor the health in the concrete structures the authors have been developing the prediction and identification models that also reduce the manual labour and errors that occur due to human miscalculations, thus the automated machine learning based models are being adopted by eradicating human intervention in the crack detection. Also, to make the models highly efficient and accurate the reliable models like CNN, ANN, ResNet and Vision Transformers (ViT) have been adopted by the researchers more, based on their necessity and research objectives.

2.3. Transformers

GEHRI *et al*. [6] used digital image correlation technique to identify and detect cracks in the images through automated detection model where the model achieved higher accuracy in detecting even the complex cracks than micro and macro cracks in the concrete structures. The authors used kinetic measurement technique and achieved higher accuracy. However, the kinetic measurement method had biased outcomes in closely-spaced cracks. Crack detection models that have bias in complex crack detection will result with irregular outcomes and inconsistent results that could affect the research. Henceforth the authors [3] proposed a model later on with 'slope surface' crack detection from images using deep learning. The authors found that vision transformer model achieved 94% accuracy whereas the other models (LeNet, AlexNet, InceptionA, MobileNet, ResNet and InceptionE) achieved higher accuracy as 99% each respectively. So, in the year 2021, authors [31] proposed a vision transformer model (SOTA) through visual interpretation technique that uses CNN architecture. The authors used CNN architecture model and ViT model with Chinese and German asphalt as samples. The accuracy obtained by the ViT model with German samples was 99% and in the Chinese samples was 91%. The authors thus proved that ViT model achieves higher accuracy too with lesser cost and time for processing the data.

Similarly, the authors [4] also proposed visual interpretation through deep learning and authors [7] proposed a deep learning model for slope surface-based crack detection. Both studies compared CNN and ViT architectures and found that CNN proved to be higher in accuracy score than ViT model. However, they also insisted that, cost, time, labour and resources were also high in the CNN model along with its stacked-up layers. Henceforth they concluded and claimed that ViT models are better in crack detection for small budget researches with better accuracy and CNN models are costlier but highly efficient with accurate outcomes. YU *et al*. [32] examined the damages in building structures using the deep learning CNN model and authors [33] examined vision-based crack detection in concrete structures using the hybrid CNN model approach. Both studies concluded that CNN model is accurate in crack detection. Later, the study by [34] utilized the optimization algorithm (improved bird-swarm algorithm: IBSA) in CNN model and found it to be significantly accurate and precise with minimal loss than earlier approaches. Similarly in another study authors [35] examined the concrete structures with CNN model with enhanced chicken-swarm algorithm (ECSA) and found that, CNN models achieve higher accuracy and performance in detecting cracks. Thus it is comprehensible that, utilizing an optimization algorithm with CNN model certainly increases the accuracy than a hybrid or a normal CNN model architecture.

Hence to prove these researches and how the samples might differ in the ViT model crack detection and CNN crack detection models, the current study aims at comparing the outcomes of the CNN and ViT models based on accuracy, computing cost and processing time.

3. PROPOSED METHOD

The proposed method for the developed research includes different processing phases in this research. The flow of the research (refer to Figure 1) is:

- Pre-processing the images of the concrete structures,
- Evaluating, classifying and categorizing the images,
- Segmenting the cracks and non-cracks images and storing in separate folders and
- Measurement of cracks and comparison done by both models.

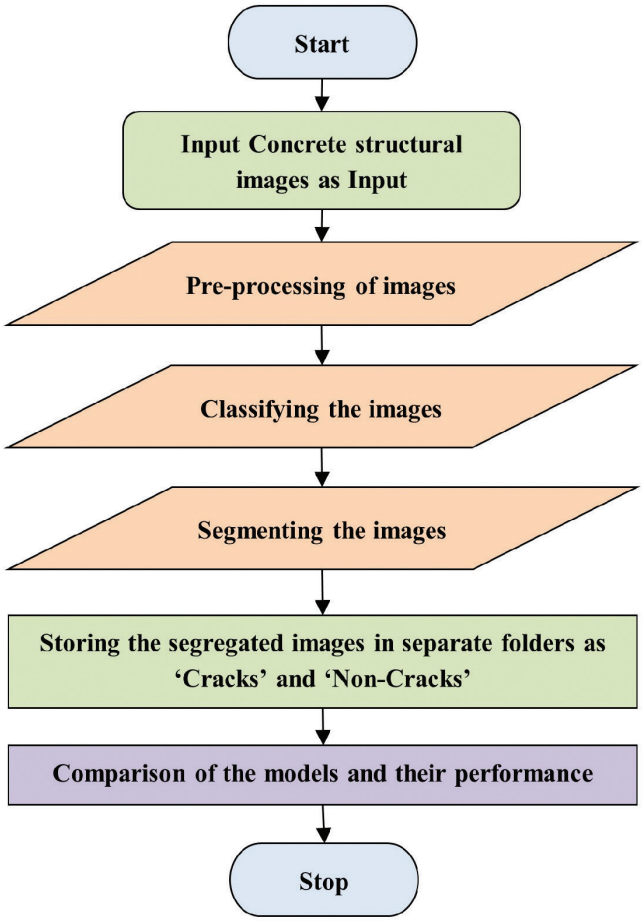

Figure 1: Flow-chart of the VGG16-Net model developed.

The model developed includes stages where the concrete structural image is inputted to the model. Then the image is later on pre-processed, classified as cracks and non-cracks and the outcomes/results are segregated and the final output obtained is stored and the models are compared for the accuracy as the metric evaluation.

3.1. Architecture of the proposed model

The model proposed is the VGG16-Net model for the detection of the concrete structural health. The images are obtained, pre-processed and then the model is applied on the datasets.

The developed model has 7 stages. Initially the input as image (i.e. concrete structure) is passed to the model as the first stage. In the second stage, 2 × convolutional layers are stacked up with 224 × 224 × 64 size. In the following stage three, a single max-pool layer and 2 × convolutional layers is stacked up with the size of 112 × 112 × 128. In the fourth stage, one max-pool layer and 3 × convolutional layers are stacked up with 56 × 56 × 256 as the size. Followed by the same set of layers (1 × max-pooling and 3 × convolutional layers) as above, where, 28 × 28 × 512 as fifth stage layer size and 14 × 14 × 512 as layer size as sixth stage is stacked up in the model.

In the final seventh stage, a max pooling layer with 7 × 7 × 512 with 500 batches as size for fully-connected layer and 500 batches as size for dropout layer with final layer of Softmax with 2 as batch size is stacked up (refer to Figure 2).

3.2. Google transformer: adopted architecture

The ViT model adopted is the model developed by the authors [36]. The model has 7 layers. The first layer includes the embedding layer where the 3 × convolutional 2D of size 768 that has kernel size of 16 × 16 and stride of 16 × 16. The second layer is the encoder layer that includes the 7 × ViT layers that has ViT layers as self-attention, self-output, intermediate, output and 2 × LayerNorm of 768. Followed by the dropout layer as third layer where p = 0.1 and next fourth layer a linear layer of 768 batch size of output featuring 256. The same is repeated in the following layers of fifth and sixth layer where the dropout layer with p = 0.2 and

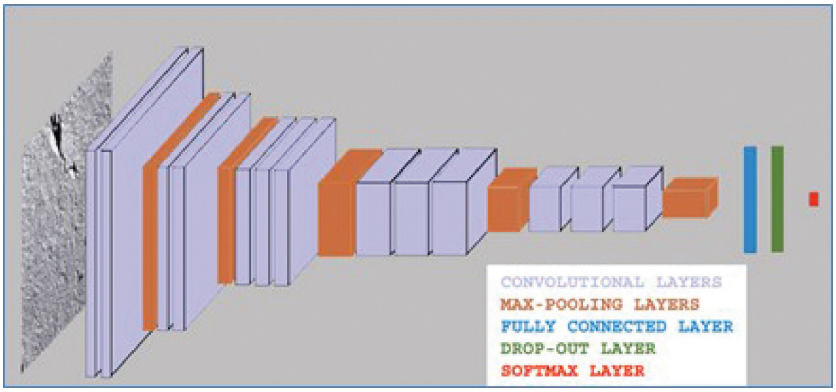

Figure 2: Architecture of the VGG16-Net model adopted.

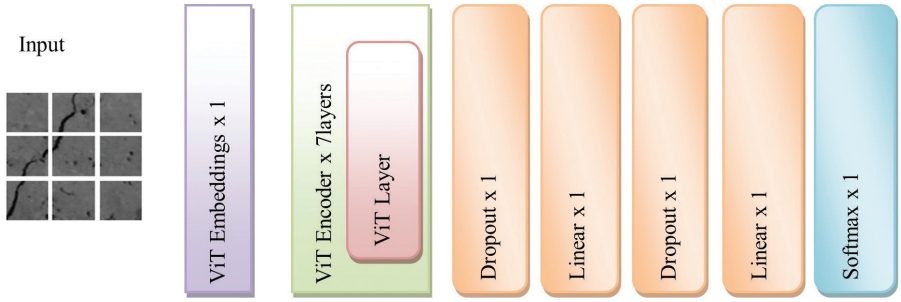

Figure 3: Architecture of the vision transformer adopted (Google model).

next a linear layer of 256 batch size of output featuring 2. The final layer is the Softmax as the seventh layer (refer to Figure 3).

The transformer model is applied on datasets where the images of the concrete structures are focused and detected for the cracks and non-cracks in the images. The detected images are then classified into separate folders with classes “cracks” and “non-cracks”.

3.3. Categorical cross-entropy (loss) function

In this research the loss is estimated through computing the ‘Softmax’ function for the developed model. The Softmax loss function is considered neither as activation nor as a loss function. It is rather identified as a Cross-Entropy function in estimating the loss for detection and image classification models. However, researchers also adopt this technique as activation and loss estimation function according to their necessity. Here the formula used is:

$$\delta(\bar{v})_a = \frac{e^{a_x}}{\Sigma_{y=1}^{C} e^{a_y}} \tag{1}$$

Where:

δ = Softmax, $\bar{v}$ = Input vector, C = total classes in multi-class classifier, e^{a_x} = input vector’s standard exponential function and e^{a_y} = output vector’s standard exponential function.

3.4. Convolutional neural network (CNN)

The convolutional neural network (CNN) in Pytorch is used by the researchers for small datasets and also for large datasets. In this research the CNN is used for the developed model where the pre-trained weights are used. For pre-processing the images (size, resolution, colour, brightness, contrast and dimension) the researcher has set the size as 224×224 pixels for the images. Once the images are pre-processed the research adopts the CNN layers convolutional layers, max pooling layers, Softmax layer, dropout layer and fully connected layer. In the initial stage the input is passed to the convolutional layers and max pooling layers. Once the image is passed

through the layers, it is then passed on to the dropout layer and fully connected layer. Finally, the Softmax layer is connected and the outcome is processed and segregated post classification as "cracks" and "non-cracks".

3.5. Adam optimizer as adopted algorithm

Adam optimizer (AOA) is mostly used by the researchers in the deep-learning based identification and prediction 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 lesser processes than other algorithms and thus preferred by researchers majorly. This research being the comparative study of ViT model and VGG16-Net model the researcher had adopted the Adam optimization for comparing the outcomes. The pseudo-code for the adopted algorithm is:

Adam Optimizer Algorithm: Input

Step 1: Initialize the process, $\boldsymbol{a_b = 0}$ and $\boldsymbol{x_1}$

Step 2: For, n = 1, ..., N, do

$$\boldsymbol{a_n = S_{1,n} a_{n-1}} + (1 - \boldsymbol{\rho_{1,n}})\boldsymbol{g_n}$$

$$\hat{v}_n = h_n(\boldsymbol{g_1}, \boldsymbol{g_2}, \ldots, \boldsymbol{g_n})$$

$$\boldsymbol{x_{n+1}} = \boldsymbol{x_n} - \frac{\boldsymbol{m_n b_n}}{\sqrt{\hat{v}_n}}$$

Step 3: End the process.

By using the AOA-algorithm the researcher uses the model to identify and classify the concrete structures and classify them as cracked and non-cracked structures.

4. EXPERIMENT AND RESULTS

Firstly, the experimentation for the developed model is carried out through applying the VGG16-Net model on datasets. The model is tested trained and the outcomes obtained are stored under the classified labels. Secondly, the experimentation is carried out for ViT model and the outcomes are stored for comparison against the VGG16 model. Through the experimentation the outcomes from the models are evaluated and compared for better performing model with higher accuracy.

4.1. Setting-up the experiment

The experiment is conducted on the datasets acquired from the kaggle resource of concrete structures where the developed model is aimed to detect the cracks and non-cracks and label the outcomes as classified images in separate folders. The developed 'VGG16-Net model' includes the Adam optimization algorithm. For the model the weights that are pre-trained from the Google patch (google/vit-base-patch16-224-in21k) has been used. The library used here is the "Hugging Face" transformers.

4.2. Datasets

The images of the cracked and non-cracked cement structures as the datasets are acquired from the resource "Kaggle" by [37]. The total datasets accumulate to images of 20 thousand and more concrete structures from buildings, bridges and pavements. Among which the research has used the 70% (14000:14000 for cracks and non-cracks as image classification) for training and the rest 30% datasets (6000:6000 for both cracks and non-cracks image classification) for testing and validation. Thus, the split-ratio for the obtained dataset is categorized as 70:20:10 for training, testing and validation respectively.

4.3. Training and testing

For training, the images are first pre-processed and then the first model (VGG16-Net) is applied on the dataset. The outcomes are observed until the model's accuracy is higher and constant. Until the model obtains high accuracy the model is trained and the final model with higher accuracy is retained. The final VGG16-Net model is then compared against the ViT model.

The VGG16-Net model is trained with 70% datasets and the loss is evaluated for the model where the cross-entropy is used. The loss functions are:

4.3.1. Loss function

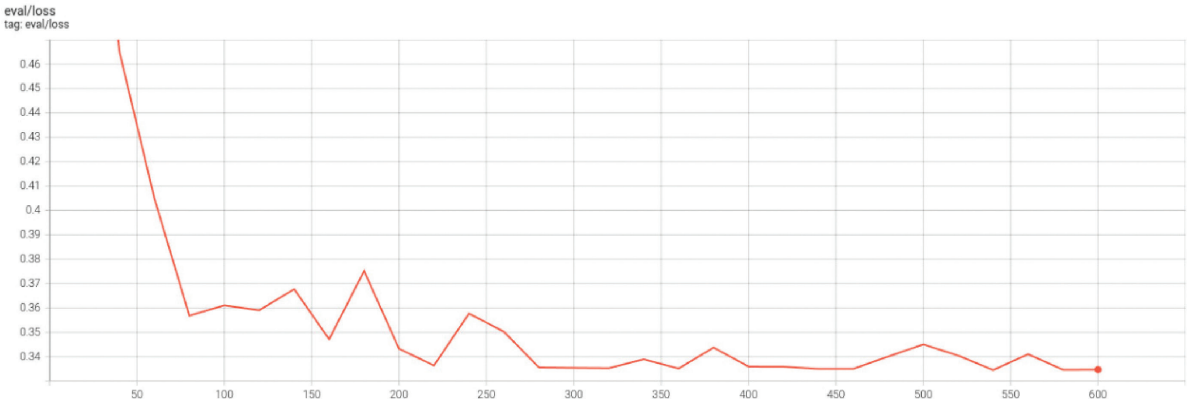

Figure 4: Evaluation loss.

Inference:

The evaluation loss (refer to Figure 4) was observed as high (0.47) at the initial epoch training. Later around the 80th epoch the loss reduced from 0.47 to 0.35. However, the loss fluctuated from 125th epoch till 275th epoch and remained reducing. The loss attained the value 0.34 at 575th epoch and remained same until 600th.

Figure 5: Training loss.

The training loss (refer to Figure 5) of the model is evaluated and the loss is observed initially at the 47th epoch at 0.59. The loss reduced from 0.58 to 0.38 at 75th epoch. Later on, the loss fluctuated and at epoch 380th the loss reduced at 0.36. Finally at the epoch 600th the loss value reduced to 0.356.

4.4. Performance metric evaluation

The evaluation metrics for the developed VGG16-Net model is calculated through the metric evaluation technique.

The techniques are:

4.4.1. F1-score

The F1-score for the model is calculated through:

$$F1\text{-}Score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \tag{2}$$

From Figure 6 the highest F1-Score has been observed as 0.985 at the 600th epoch.

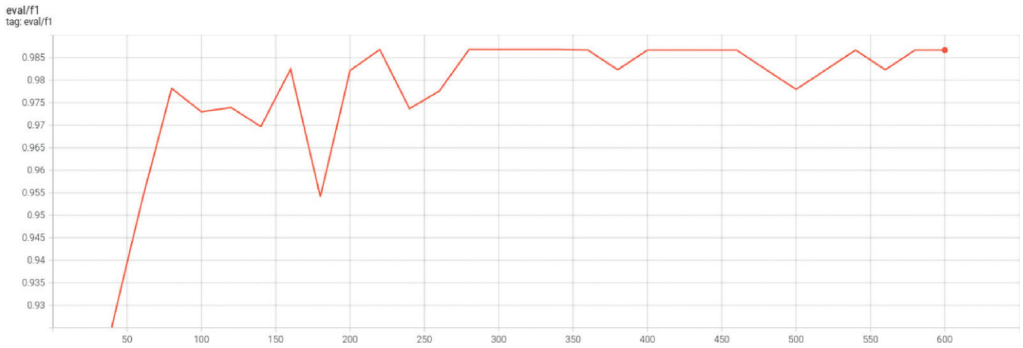

Figure 6: F1-score of VGG16-Net model.

4.4.2. Recall

The Recall for the model is calculated through:

$$Recall = \frac{TruPos}{TruPos + FalNeg} \tag{3}$$

Where: TruPos denotes true-positives, TruFal denotes true-negatives, FalNeg denotes false-negatives and FalPos denotes false-positives.

From Figure 7 the highest recall has been observed as 0.991 from 360th to 600th epoch.

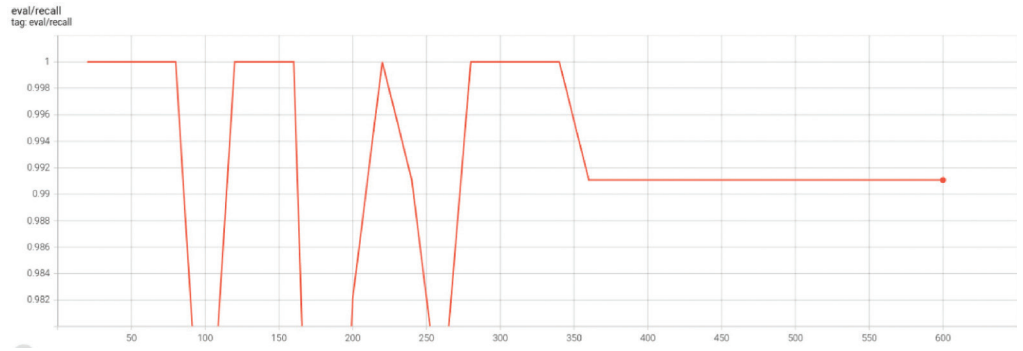

Figure 7: Recall of VGG16-Net model.

4.4.3. Precision

The Precision for the model is calculated through:

$$Precision = \frac{TruPos}{TruPos + FalPos} \tag{4}$$

From Figure 8 the highest precision has been observed as 0.98 at the 600th epoch.

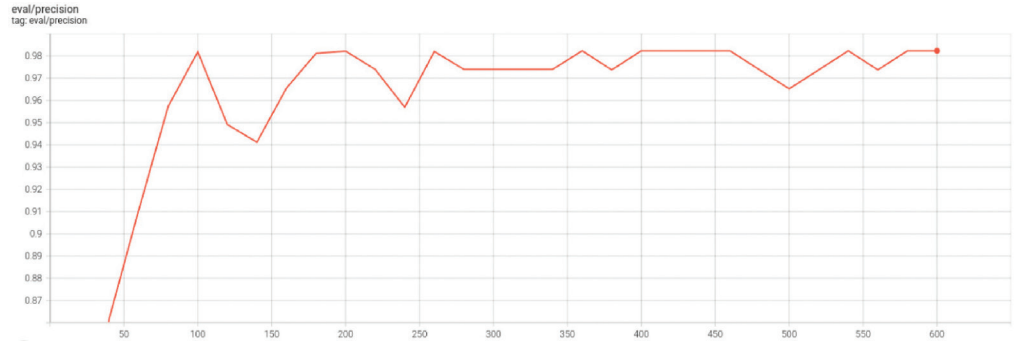

Figure 8: Precision of VGG16-Net model.

4.4.4. Accuracy

The Accuracy for the model is calculated through:

$$Accuracy = \frac{TruPos + TruNeg}{TruPos + FalPos + TruNeg + FalNeg} \tag{5}$$

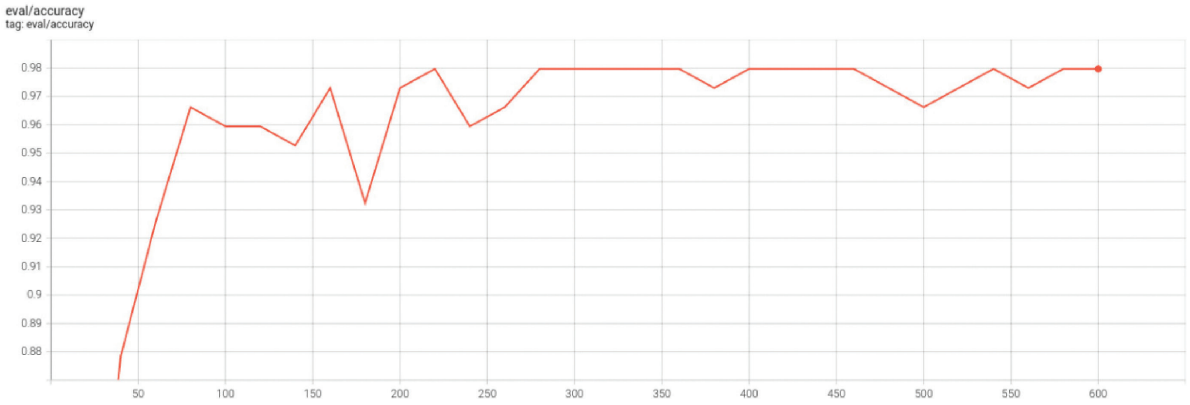

Figure 9: Accuracy of VGG16-Net model.

From Figure 9 the highest accuracy has been observed as 0.98 at the 600th epoch.

The final training and testing of the model was done with 1261 samples for training and 239 for validation with 32 batch sizes and 165 samples for testing with 16 batch sizes. The epoch size was 10.

The loss from Table 1 could be observed where at the 10th epoch the accuracy increased as the losses decreased for both training and validation.

4.5. Experimenting results

The outcomes of the VGG16 model are classified as two classes, where the results obtained are segregated and categorized as 'Cracks' and 'Non-cracks':

The prediction was made for the VGG16 model with the concrete images and the following outcomes are obtained where the label cracks are applied on the classified imagespre-storing the results:

The prediction was made for the VGG16 model with the concrete images and the following outcomes are obtained where the label Non-cracks are applied on the classified images prior storing the results:

From the results from Tables 2 and 3 it can be observed that, the prediction made by the VGG16 model was high where among the 10 sample datasets only one was found to be predicted wrongly.

Table 1: Epoch table for loss evaluation.

	TIME/STEP (IN SECOND AND MILLISECONDS)	TRAINING-LOSS	TRAINING-ACCURACY	VALIDATION-LOSS	VALIDATION-ACCURACY
1	26s 675ms	0.5381	0.7697	0.5768	0.6875
2	30s 767ms	0.3971	0.8039	0.4241	0.7500
3	23s 571ms	0.3396	0.8446	0.3104	0.8616
4	24s 605ms	0.2256	0.8999	0.3926	0.8616
5	22s 562ms	0.2266	0.9056	0.2000	0.9152
6	22s 561ms	0.1340	0.9504	0.1427	0.9330
7	22s 571ms	0.1099	0.9536	0.2263	0.8973
8	22s 562ms	0.1422	0.9479	0.1233	0.9598
9	22s 563ms	0.0975	0.9642	0.0845	0.9598
10	22s 560ms	0.0747	0.9715	0.1093	0.9598

Table 2: Results of the VGG model classification – Crack.

MODEL OUTPUT	PREDICTED OUTCOME	ACTUAL OUTCOME
	Crack	Crack
	Crack	Crack
	Crack	Crack
	Non-crack	Crack
	Crack	Crack

Table 3: Results of the VGG model classification – Non-crack.

MODEL OUTPUT	PREDICTED OUTCOME	ACTUAL OUTCOME
	Non-crack	Non-crack
	Non-crack	Non-crack
	Non-crack	Non-crack
	Non-crack	Non-crack
	Non-crack	Non-crack

4.6. Comparative analysis of the models developed

The performances of both models are evaluated through metric evaluation techniques. The outcomes are (refer to Table 4):

Table 4: Comparative analysis of the models.

MODEL	PRECISION	ACCURACY	RECALL	F1-SCORE
ViT	97.30%	92.72%	92.50%	94.87%
VGG16-Net	98.0%	98.0%	99.10%	98.50%

The validation accuracy as the metric for evaluating the performance of the models 'ViT' and 'VGG16-Net' is measured (refer to Table 2). The ViT model achieved 96.6% and the VGG16-Net model achieved 95.98% validation accuracy.

The training of the VGG16-Net model was carried out by increasing the epochs from 10 to 30 with 16 batch size and learning rate 0.1.

From Table 5 it can be observed that the loss has been reduced from 0.65 to 0.35 from epoch 1 to epoch 30.

Table 5: Epoch table – performance evaluation.

EPOCH	TRAINING LOSS	VALIDATION LOSS	ACCURACY	F1	PRECISION	RECALL
1	0.651200	0.610249	0.756757	0.861538	0.756757	1.000000
2	0.515300	0.465112	0.878378	0.925620	0.861538	1.000000
3	0.414400	0.404916	0.925676	0.953191	0.910569	1.000000
4	0.371500	0.356831	0.966216	0.978166	0.957265	1.000000
5	0.373800	0.361056	0.959459	0.972973	0.981818	0.964286
6	0.388200	0.359046	0.959459	0.973913	0.949153	1.000000
7	0.389400	0.367694	0.952703	0.969697	0.941176	1.000000
8	0.392800	0.347127	0.972973	0.982456	0.965517	1.000000
9	0.398400	0.375145	0.932432	0.954128	0.981132	0.928571
10	0.371300	0.343238	0.972973	0.982143	0.982143	0.982143
11	0.368900	0.336343	0.979730	0.986784	0.973913	1.000000
12	0.360500	0.357701	0.959459	0.973684	0.956897	0.991071
13	0.379500	0.350253	0.966216	0.977578	0.981982	0.973214
14	0.363100	0.335580	0.979730	0.986784	0.973913	1.000000
15	0.379600	0.335415	0.979730	0.986784	0.973913	1.000000
16	0.382500	0.335282	0.979730	0.986784	0.973913	1.000000
17	0.375600	0.338903	0.979730	0.986784	0.973913	1.000000
18	0.365300	0.335134	0.979730	0.986667	0.982301	0.991071
19	0.353200	0.343718	0.972973	0.982301	0.973684	0.991071
20	0.363100	0.335931	0.979730	0.986667	0.982301	0.991071
21	0.357300	0.335889	0.979730	0.986667	0.982301	0.991071
22	0.362300	0.335003	0.979730	0.986667	0.982301	0.991071
23	0.363900	0.335021	0.979730	0.986667	0.982301	0.991071
24	0.358000	0.340056	0.972973	0.982301	0.973684	0.991071
25	0.358400	0.345010	0.966216	0.977974	0.965217	0.991071
26	0.353700	0.340439	0.972973	0.982301	0.973684	0.991071
27	0.353300	0.334514	0.979730	0.986667	0.982301	0.991071
28	0.358000	0.340995	0.972973	0.982301	0.973684	0.991071
29	0.363700	0.334628	0.979730	0.986667	0.982301	0.991071
30	0.356100	0.334709	0.979730	0.986667	0.982301	0.991071

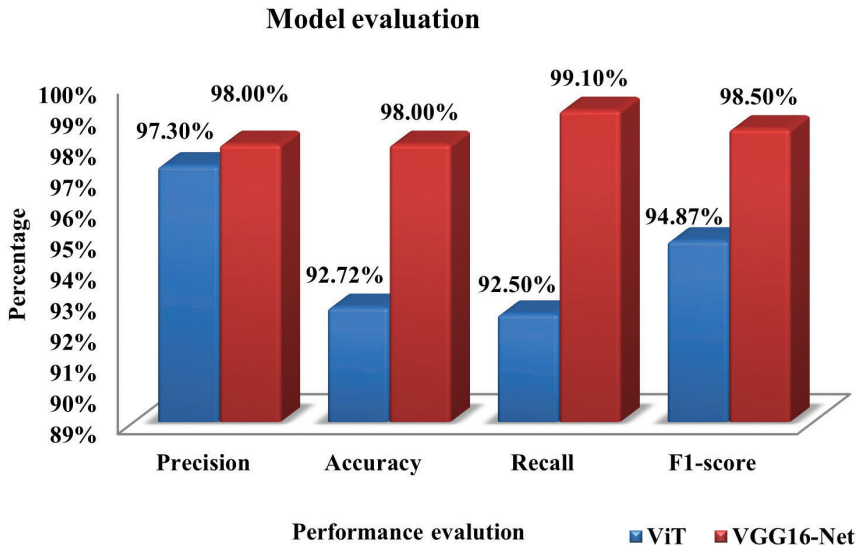

Figure 10: Comparative analysis of ViT & VGG16.

Figure 10 illustrates the comparison of the models ViT (vision transformer) and VGG16-Net (CNN). From the observation it is understood that, the outcomes obtained from the VGG-16 model for identifying and detecting the cracked and non-cracked concrete structures are better than the transformer model. The evaluation metrics shows that accuracy, recall and f1-scores are higher in VGG16+CNN model than ViT model; whereas the precision of ViT model is somewhat similar to the VGG16 model. This finding is similar to conclusions of the studies by authors [3,6,12,13] where CNN models are better in accuracy than ViT models.

From Figure 11 it is apparent that the developed VGG16 model acquired more accuracy than existing convolutional neural networking models like AlexNet, GoogleNet, InceptionNet and ResNet.

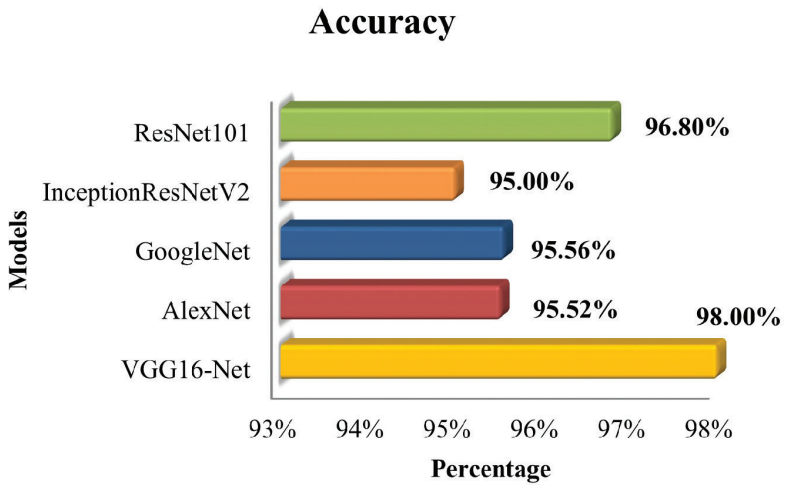

Figure 11: Comparison of few different CNN models.

Table 6: Accuracy comparison of CNN models.

AUTHOR & YEAR	MODEL	ACCURACY
Proposed model, 2022	VGG16-Net	98.00%
[38]	AlexNet	95.52%
[38]	GoogleNet	95.56%
[39]	InceptionResNetV2	95.00%
[40]	ResNet101	96.80%

5. FINDINGS

The findings from the evaluated outcomes are presented in Table 6.

- The scores of the VGG16-Net model are: precision at 98.0%, accuracy at 98.0%, f1-score at 98.50% and recall at 99.10%.
- The scores of the ViT model are: precision at 97.30%, accuracy at 92.72%, recall at 92.50% and f1-score at 94.87%.
- The validation accuracy of the ViT model (96.6%) and the VGG16-Net model (95.98%) was found to be 96%.

Hence it is found that when the computing cost, scheduling time, processing time, constructing stacked-up layers and other resources are considered ViT crack detection model is better however when the accuracy, precision, recall and f1-score are considered the CNN crack detection model is better.

6. DISCUSSION AND CONCLUSION

The existing researches on crack detection models were made by examining the detection models processing speed, processing time and how accurate they are. However, the lack of comparative studies on how the novel and traditional models differ had not been done in machine learning models for the crack detection models in concrete structures. The current study aimed at applying the VGG model and ViT model on datasets of concrete images through CNN architecture in VGG. The images were acquired from kaggle as resource. Datasets acquired were more-than 20 thousand concrete images of buildings and pavements. The secondary datasets were journals, e-journals, articles, research papers and studies on concrete structural health through machine learning.

The cross-entropy was used as loss function to evaluate and examine the training and validation loss of the model. The Adam optimization as the model's algorithm was utilized in the developed VGG model neural layers to increase the accuracy with minimal loss through fine tuning the hyper-parameters learning rate and epochs. To avoid overfitting (too many) and underfitting (too few) of data, the epochs were retained to as 30 with the learning rate of 0.1. The training and testing and validation of datasets were split in 70:20:10 ratio and the outcomes are segregated as two classes, namely, "cracks" and "non-cracks". The images that are predicted with cracks are classified and stored under the folder "cracks" and the images that are classified and identified as "non-cracks" are stored under the non-cracks folder. The outcomes of the prediction made by the ViT model and the VGG16 model are compared. Both models (VGG and ViT) have acquired 96% as validation accuracy. The research examined the performance metrics, where the VGG model acquired higher accuracy (98%) than the ViT model. However, since a VGG model is stacked up with the CNN layers, the computing time and other resource costs are considerably higher than vision transformers. Thus, it's safer to conclude that, when a researcher aims at a detection model with better performance and higher accuracy it is considerable to utilize the traditional models with CNN, like VGG16 model, AlexNet, LeNet, ResNet model and other models. Contrarily if a researcher concentrates on lesser resources and rapid time for processing it is better to utilize the ViT model to minimize the resource expenses.

Through the findings the study concludes that, by utilizing the vision transformer model to detect and classify the crack identification images in structural engineering saves costs and processing time. Contrarily, the CNN based models with VGG16 architecture is simpler, customizable, more matured to implement and also to be trained than vision transformers that utilizes image-net. ViT splits-up the images as several visual tokens whereas the CNN utilizes the pixel arrays. The differences in accuracy of the models are marginal, but when the benefits and advantages are weighed-in for structural health monitoring projects, the VGG16 + CNN architecture is better with accuracy and performance than vision transformers.

Future enhancements: The current study developed a VGG16 model and compared the results with the ViT model. However, on the future the same could be extended where instead of VGG16 model other models like ResNet, LeNet, AlexNet and other architectures without CNN layers could be used for comparing the outcomes with the ViT model. Similarly, the future research could focus on models of crack prediction in concrete structures by examining their precision, accuracy, recall and f1-score to find the best model in predicting and detecting cracks. In future, more CNN based architectures will be analysed and studied for performance and accuracy. Different evaluation metrics will 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) will also be focused to find the better NN models in crack detection, than CNN and vision transformers.

Limitations: The study aims at comparing the CNN based VGG16 detection model against the ViT model and thus other architectures are neglected. Since the study utilizes the resources from kaggle, the UAV (unmanned aerial-vehicles) based researches are not studied. Thus, the research exclusively focuses on the purpose and limits the focus on vision transformer and VGG16 models alone.

7. REFERENCES

[1] YUAN, G., LI, J., MENG, X., *et al*., "CurSeg: a pavement crack detector based on a deep hierarchical feature learning segmentation framework", *IET Intelligent Transport Systems*, v. 16, n. 6, pp. 782–799, 2022. doi: http://dx.doi.org/10.1049/itr2.12173.

[2] PARAMANANDHAM, N., KOPPAD, D., ANBALAGAN, S., "Vision based crack detection in concrete structures using cutting-edge deep learning techniques", *TS. Traitement du Signal*, v. 39, n. 2, pp. 485–492, 2022. doi: http://dx.doi.org/10.18280/ts.390210.

[3] CHEN, Y., GU, X., LIU, Z., *et al*., "A fast inference vision transformer for automatic pavement image classification and its visual interpretation method", *Remote Sensing*, v. 14, n. 8, pp. 1877, 2022. doi: http://dx.doi.org/10.3390/rs14081877.

[4] XU, Z., GUAN, H., KANG, J., *et al*., "Pavement crack detection from CCD images with a locally enhanced transformer network", *International Journal of Applied Earth Observation and Geoinformation*, v. 110, pp. 102825, 2022. doi: http://dx.doi.org/10.1016/j.jag.2022.102825.

[5] SILVA, W.R.L.D., LUCENA, D.S.D., "Concrete cracks detection using deep learning image classification", *Proceedings*, v. 2, n. 489, pp. 1–6, 2018.

[6] GEHRI, N., MATA-FALCON, J., KAUFMANN, W., "Automated crack detection and measurement based on digital image correlation", *Construction & Building Materials*, v. 256, pp. 119383, 2020. doi: http://dx.doi.org/10.1016/j.conbuildmat.2020.119383.

[7] YANG, Y., MEI, G., "Deep transfer learning approach for identifying slope surface cracks", *Applied Sciences*, v. 11, n. 23, pp. 11193, 2021. doi: http://dx.doi.org/10.3390/app112311193.

[8] HE, K., ZHANG, X., REN, S., *et al*., "Deep residual learning for image recognition", *IEEE Proceedings on Computer Vision and Pattern Recognition (CVPR)*, v. 2021, pp. 770–778. 2016.

[9] KRIZHEVSKY, A., SUTSKEVER, I., HINTON, G.E., "ImageNet classification with deep convolutional neural networks", *Communications of the ACM*, v. 60, n. 6, pp. 84–90, 2017. doi: http://dx.doi.org/10.1145/3065386.

[10] SIMONYAN, K., ZISSERMAN, A., "Very deep convolutional networks for large-scale image recognition", In: *Proceedings of the 3rd International Conference of Learning Representations (ICLR)*, pp. 1–12, 2015.

[11] LECHUN, Y., BOTTOU, L., BENGIO, Y., *et al*., "Gradient - based Learning Applied to Document Recognition", *IEEE Proceedings,* v. 86, pp. 2278–2323, 1998.

[12] SZEGEDY, C., LIU, W., JIA, Y., *et al*., "Going deeper with convolutions", *IEEE Proceedings of Computer Vision and Pattern Recognition (CPVR)*, v. 2015, pp. 1–9. 2015.

[13] SZEGEDY, C., VANHOUCKE, V., LOFFE, S., *et al*., "Rethinking the inception architecture for computer vision", *IEEE Proceedings: Computer Vision and Pattern Recognition (CVPR)*, v. 2016, pp. 2818–2826. 2016.

[14] HOWARD, A.G., ZHU, M., CHEN, B., *et al*., "*MobileNets: efficient convolutional neural networks for mobile vision applications*, 2017, https://www.semanticscholar.org/reader/3647d6d0f151dc05626449ee09cc7bce55be497e, accessed in August, 2020.

[15] HUANG, G., LIU, Z., VAN DER MAATEN, L., *et al*., "Densely connected convolutional networks", *IEEE Proceedings on Computer Vision and Pattern Recognition (CVPR)*, v. 2017, pp. 2261–2269. 2017.

[16] FANG, W., LUO, H., XU, S., *et al*., "Automated text classification of near-misses from safety reports: an improved deep learning approach", *Advanced Engineering Informatics*, v. 44, pp. 101060, 2020. doi: http://dx.doi.org/10.1016/j.aei.2020.101060.

[17] ISLAM, M.M.M., KIM, J.-M., "Vision-based autonomous crack detection of concrete structures using a fully convolutional encoder–decoder network", *Sensors*, v. 19, n. 19, pp. 4251, 2019. doi: http://dx.doi.org/10.3390/s19194251. PubMed PMID: 31574963.

[18] KIM, J.J., KIM, A.-R., LEE, S.-W., "Artificial neural network-based automated crack detection and analysis for the inspection of concrete structures", *Applied Sciences*, v. 10, n. 22, pp. 8105, 2020. doi: http://dx.doi.org/10.3390/app10228105.

[19] ARBAOUI, A., OUAHABI, A., JACQUES, S., *et al*., "Concrete cracks detection and monitoring using deep learning-based multiresolution analysis", *Electronics*, v. 10, n. 15, pp. 1772, 2021. doi: http://dx.doi.org/10.3390/electronics10151772.

[20] DORAFSHAN, S., THOMAS, R.J., MAGUIRE, M., "SDNET2018: an annotated image dataset for non-contact concrete crack detection using deep convolutional neural networks", *Data in Brief*, v. 21, pp. 1664–1668, 2018. doi: http://dx.doi.org/10.1016/j.dib.2018.11.015. PubMed PMID: 30505897.

[21] ADAM, E.E.B., SATHESH, A., "Construction of accurate crack identification on concrete structure using hybrid deep learning approach", *Journal of Innovative Image Processing*, v. 3, n. 2, pp. 85–99, 2021. doi: http://dx.doi.org/10.36548/jiip.2021.2.002.

[22] PARK, M.J., KIM, J., JEONG, S., *et al*., "Machine learning-based concrete crack depth prediction using thermal images taken under daylight conditions", *Remote Sensing*, v. 14, n. 9, pp. 2151, 2022. doi: http://dx.doi.org/10.3390/rs14092151.

[23] ALI, L., ALNAJJAR, F., JASSMI, H.A., *et al*., "Performance evaluation of deep CNN - based crack detection and localization techniques for concrete structures", *Sensors*, v. 21, n. 5, pp. 1688, 2021. doi: http://dx.doi.org/10.3390/s21051688.

[24] COCA, L.-G., CUSMULIUC, C.G., IFTENE, A., "Automatic tarmac crack identification application", *Procedia Computer Science*, v. 192, pp. 478–486, 2021. doi: http://dx.doi.org/10.1016/j.procs.2021.08.049.

[25] HALLEE, M.J., NAPOLITANO, R.K., REINHART, W.F., *et al*., "Crack detection in images of masonry using CNNs", *Sensors*, v. 21, n. 14, pp. 4929, 2021. doi: http://dx.doi.org/10.3390/s21144929. PubMed PMID: 34300668.

[26] REHUMAAN, S.F.K., "Modelling of cracked concrete and identification of design parameters using static non-linear analysis", *Civil Engineering and Architecture*, v. 10, n. 2, pp. 584–599, 2022. doi: http://dx.doi.org/10.13189/cea.2022.100216.

[27] SU, C., WANG, W., "Concrete cracks detection using convolutional neural network based on transfer learning", *Mathematical Problems in Engineering*, v. 2020, pp. 1–10, 2020.

[28] ZHANG, H., LIU, C., HO, J., *et al*., "Crack detection based on convnext and normalization", *Journal of Physics: Conference Series*, v. 2289, n. 1, pp. 012022, 2022. doi: http://dx.doi.org/10.1088/1742-6596/2289/1/012022.

[29] REN, Y., HUANG, J., HONG, Z., *et al*., "Image-based concrete crack detection in tunnels using deep fully convolutional networks", *Construction & Building Materials*, v. 234, pp. 117367–117378, 2020. doi: http://dx.doi.org/10.1016/j.conbuildmat.2019.117367.

[30] SATHYA, K., SANGAVI, D., SRIDHARSHINI, P., *et al*., "Improved image based super resolution and concrete crack prediction using pre-trained deep learning models", *Journal of Soft Computing in Civil Engineering*, v. 4, n. 3, pp. 40–51, 2020.

[31] DOSOVITSKIY, A., BEYER, L., KOLESNIKOV, A., *et al*., "An image worth 16*16 words: transformers for image recognition at scale", *ICLR Conference*, v. 2021, pp. 1–22, 2021.

[32] YU, Y., WANG, C., GU, X., *et al*., "A novel deep learning-based method for damage identification of smart building structures", *Structural Health Monitoring*, v. 18, n. 1, pp. 143–163, 2019. doi: http://dx.doi.org/10.1177/1475921718804132.

[33] YU, Y., SAMALI, B., RASHIDI, M., *et al*., "Vision-based concrete crack detection using a hybrid framework considering noise effect", *Journal of Building Engineering*, v. 61, pp. 105246, 2022. doi: http://dx.doi.org/10.1016/j.jobe.2022.105246.

[34] YU, Y., LIANG, S., SAMALI, B., *et al*., "Torsional capacity evaluation of RC beams using an improved bird swarm algorithm optimised 2D convolutional neural network", *Engineering Structures*, v. 273, pp. 115066, 2022. doi: http://dx.doi.org/10.1016/j.engstruct.2022.115066.

[35] YU, Y., RASHIDI, M., SAMALI, B., *et al*., "Crack detection of concrete structures using deep convolutional neural networks optimized by enhanced chicken swarm algorithm", *Structural Health Monitoring*, v. 21, n. 5, pp. 2244–2263, 2022. doi: http://dx.doi.org/10.1177/14759217211053546.

[36] VASWANI, A., SHAZEER, N., PARMAR, N., *et al*., "Attention is all you need", In: *Proceedings of the 31st International Conference on Neural Information Processing Systems,* v. 30, pp. 5998–6008, 2017.

[37] SRIKAR, B., "Concrete crack images for classification", 2019, https://www.kaggle.com/datasets/thesighsrikar/concrete-crack-images-for-classification, accessed in August, 2020.

[38] SHARMA, N., DHIR, R., RANI, R., "Crack detection in concretes using transfer learning", *Advances in Mathematics: Scientific Journal*, v. 9, n. 6, pp. 3895–3906, 2020.

[39] RAJADURAI, R.-S., KANG, S.-T., "Automated vision-based crack detection on concrete surfaces using deep learning", *Applied Sciences*, v. 11, pp. 5229, 2021. doi: http://dx.doi.org/10.3390/app11115229.

[40] MENG, X., "Concrete crack detection algorithm based on deep residual neural networks", *Scientific Programming*, v. 2021, pp. 3137083, 2021. doi: http://dx.doi.org/10.1155/2021/3137083.