بسم الله الرحمن الرحيم

King Fahd university of Petroleum & Minerals

Deanship of Research

جامعة الملك فهد للبترول والمعادن

عمادة البحوث

Undergraduate Research Office (URO)

Guided Research Program - Term 241

Final Report

To be prepared by the student not the advisor

Student's Name	Abdulrahman Ammar	Advisor' Name	Hussain Mohammad Al-Qahtani
ID	202183170	Research Center	Department of Computing and Mathematics
Department	Computation and Mathematics	Research Topic	Deep Learning-Based Crack Detection in Pipes Using CNN, ResNet, and ViT Models
Level	Junior		

Date:
16/02/2025

H. Qahtani

Student's Signature

Advisor's Signature

Deep Learning-Based Crack Detection in Pipes Using CNN, ResNet, and ViT Models

Abdulrahman Ammar
Department of Computing and Mathematics
King Fahd University of Petroleum and Minerals
s202183170@kfupm.edu.sa

***Abstract*- Detecting cracks in pipelines is a crucial task for ensuring structural integrity in industries such as oil and gas, water supply, and infrastructure maintenance. This research investigates the application of deep learning, particularly Convolutional Neural Networks (CNN) and ResNet architectures, for classifying cracked and uncracked pipes. Using a dataset of over 2000 images, we trained and evaluated models to achieve high accuracy in crack detection. The study explores the impact of data augmentation, hyperparameter tuning, and early stopping mechanisms to prevent overfitting. Experimental results indicate that the ResNet-based model outperforms the traditional CNN, achieving a test accuracy of approximately 88.5%. The research demonstrates the potential of deep learning in automating defect detection, reducing human intervention, and improving predictive maintenance strategies in pipeline monitoring. Future work will focus on integrating real-time monitoring and edge deployment for practical industrial applications.**

1. INTRODUCTION

Pipelines are crucial for transporting fluids such as oil, gas, and water across vast distances, forming the backbone of industrial infrastructure. Ensuring their structural integrity is vital for preventing failures that can lead to environmental disasters, The pipeline leak incidents have raised serious safety issues. Pipeline cracks are among the most common causes of failures, leading to leaks and even catastrophic explosions [1][2].

Traditional crack detection methods, such as manual inspection, ultrasonic testing, and infrared thermography, have limitations. Manual inspections are labor-intensive, subjective, and prone to human error, while sensor-based methods require expensive equipment and specialized expertise. Moreover, these approaches may struggle to detect small or hidden cracks, reducing their effectiveness. To address these limitations, researchers have turned to artificial intelligence (AI) and deep learning-based solutions that leverage computer vision for automatic defect detection [3].

Deep learning, particularly Convolutional Neural Networks (CNNs), has shown promising results in identifying defects in industrial applications. Residual Networks (ResNet) improve upon CNNs by addressing vanishing gradient issues, allowing for deeper networks with enhanced accuracy. More recently, Vision Transformers (ViT) have emerged as an alternative to traditional CNNs, leveraging self-attention mechanisms to capture long-range dependencies in images. Given these advancements, this research aims to compare CNN, ResNet, and ViT architecture to determine their effectiveness in classifying cracked and uncracked pipes [4][5][6].

Despite significant advancements in AI-based defect detection, several challenges remain. First, many existing studies focus on specific architectures without a direct comparison between CNN, ResNet, and ViT for crack detection. Second, deep learning models require large, high-quality datasets, and their performance varies based on architecture, training strategies, and dataset characteristics. Third, there is limited research on how ViTs perform relative to CNN-based models in industrial defect detection.

The main research questions this study seeks to answer are *"Which deep learning architecture—CNN, ResNet, or ViT—performs best in detecting cracked and uncracked pipes based on accuracy, precision, recall, and F1-score?"*

This study fills a critical gap in the literature by conducting a direct performance comparison of these architectures using a dataset of over 2000 pipeline images.

The primary objectives of this research are:

1. To develop and train three deep learning models (CNN, ResNet, and ViT) for automatic crack detection in pipeline images.
2. To evaluate and compare the performance of these models based on accuracy, precision, recall, F1-score, and confusion matrices.
3. To analyze the strengths and weaknesses of each model in detecting cracks and discuss their potential applications in real-world pipeline inspection.

This study makes the following contributions:

1. Comprehensive Evaluation: It provides an in-depth performance comparison between CNN, ResNet, and ViT architecture for crack detection in pipelines.
2. Insights into Model Suitability: The study highlights the advantages and trade-offs between CNN-based and Transformer-based models, offering insights into their industrial applicability.
3. Potential for Real-World Deployment: The findings contribute to the development of automated pipeline monitoring systems, reducing reliance on manual inspection.

2. RELATED WORK

Crack detection in industrial pipelines has been a critical research area for decades. Traditional methods, such as manual inspections and sensor-based approaches, have significant limitations in accuracy, scalability, and cost-effectiveness. With advancements in artificial intelligence (AI), deep learning has emerged as a robust alternative for automatic crack detection. This section reviews existing approaches, including traditional methods, CNN-based models, ResNet architectures, and Vision Transformers (ViT), highlighting their strengths and limitations [7].

2.1 Traditional Methods for Crack Detection

Historically, pipeline defect detection has relied on Non-Destructive Testing (NDT) techniques such as ultrasonic testing (UT), magnetic flux leakage (MFL), radiography, and infrared thermography. While effective, these methods require expensive equipment, and trained personnel, and can be time-consuming. In particular: Ultrasonic Testing (UT) is widely used but struggles with surface roughness and requires direct contact with the pipeline [8]. Infrared Thermography can detect subsurface defects based on heat variations but is sensitive to environmental conditions [9].

While these methods are reliable in controlled environments, they lack efficiency in large-scale monitoring. The advent of computer vision and AI-based approaches has revolutionized defect detection, making it faster and more scalable.

2.2 Deep Learning for Defect Detection

The rise of deep learning has significantly improved crack detection accuracy in industrial settings. Convolutional Neural Networks (CNNs) have been widely adopted due to their ability to automatically extract relevant features from images. Unlike traditional machine learning models that rely on handcrafted features, CNNs learn hierarchical representations, making them well-suited for image-based defect classification. Several studies have successfully applied CNNs to detect cracks in materials:

Adeyemi (2024) trained a CNN-based crack detection model for concrete structures, achieving over 90% accuracy [10].

2.3 CNN-Based Approaches

CNN-based models, such as AlexNet, VGG16, and EfficientNet, have been widely used in defect detection. These architectures vary in depth and complexity:

- AlexNet was among the first CNN models used for defect detection but is computationally expensive
- VGG16 provides improved accuracy but requires extensive computational resources
- EfficientNet achieves superior performance with fewer parameters but requires extensive tuning.

Recent studies have shown that CNN models, when trained on large-scale defect datasets, can achieve high accuracy but require careful regularization techniques such as dropout, batch normalization, and data augmentation to prevent overfitting [11].

2.4 ResNet and Transfer Learning

Residual Networks (ResNet) introduced skip connections, solving the vanishing gradient problem in deep networks. This architecture allows information to bypass layers, enabling deep networks to train effectively without degradation in accuracy.

Moreover, transfer learning—where models pre-trained on large datasets such as ImageNet are fine-tuned on domain-specific datasets—has been widely used in defect detection. Fine-tuning a pre-trained ResNet model significantly reduces the need for large, labeled datasets while maintaining high accuracy.

2.5 Vision Transformers (ViT)

The introduction of Vision Transformers (ViT) has challenged traditional CNN-based architecture by using self-attention mechanisms to capture long-range dependencies in images. Unlike CNNs, which relies on local feature extraction, ViTs process images as a sequence of patches, making them highly effective for texture-based defect detection.

Key advantages of ViT for defect detection include:

- Better global feature extraction than CNNs.
- Scalability to larger datasets with increased performance.
- State-of-the-art results in various image classification tasks.

However, ViTs require significantly larger datasets and computational power to train from scratch. Researchers have mitigated this issue by using pre-trained ViTs fine-tuned for defect detection. Recent studies show that ViT models achieve comparable or superior performance to CNNs for detecting structural damage in industrial components [6].

2.6 Summary of Findings from Previous Studies

Based on the literature review, key insights are as follows:

1. Traditional NDT methods are effective but costly, time-consuming, and require expert intervention.
2. CNNs offer robust feature extraction but struggle with deep architectures.
3. ResNet improves accuracy with skip connections and is commonly used with transfer learning.
4. ViTs provide superior feature extraction but require extensive computational resources.

Given these insights, this research aims to compare CNN, ResNet, and ViT for crack detection in pipelines, filling a critical gap in existing studies by directly evaluating their performance on the same dataset.

3. DATASETS REVIEW

3.1 Dataset Description

The dataset used in this study consists of 2,000 images, with 1,000 images of cracked pipes and 1,000 images of uncracked pipes. The images were collected and preprocessed for use in deep learning models. The dataset is divided into training, validation, and testing sets to ensure robust model evaluation.

3.2 Image Preprocessing

To ensure the consistency and quality of input data for deep learning models, the following preprocessing steps were applied:

1. Resizing: All images were resized to 128×128 pixels for CNN and ResNet models and 224×224 pixels for ViT.
2. Normalization: Pixel values were normalized to a [0,1] range by dividing by 255 to enhance model stability.
3. Color Channel Handling: Images were converted to RGB format to ensure compatibility with pre-trained models such as ResNet and ViT.

3.3 Data Augmentation

To improve the generalization of deep learning models and prevent overfitting, various data augmentation techniques were applied to the training dataset:

1. Rotation: Images were randomly rotated by up to 20 degrees to simulate variations in pipe positioning.
2. Horizontal and Vertical Flipping: Random flipping was applied to account for real-world variations.
3. Zooming: A zoom range of 0.2 was applied to introduce variations in the image scale.
4. Shear Transformation: Applied with a factor of 0.1 to create slight distortions.
5. Brightness Adjustment: Random brightness augmentation was used to handle lighting variations in images.

3.4 Sample Images from the Dataset

4. METHODOLOGY

Before feeding images into deep learning models, we applied preprocessing steps to standardize input data and enhance generalization.

4.1 Image Preprocessing

Resizing: Images were resized to 128×128 pixels for CNN and ResNet, and 224×224 pixels for ViT.

Normalization: Pixel values were scaled to [0,1] by dividing 255 to improve model convergence.

Color Channel Handling: All images were converted to RGB format to ensure compatibility with pre-trained models.

4.2 Data Augmentation

To improve robustness and prevent overfitting, we applied random transformations to the training dataset:

- Rotation: Images rotated by ±20° to simulate various perspectives.
- Horizontal/Vertical Flipping: Flips were applied randomly.
- Zooming: Images zoomed in by 0.2 to create size variations.
- Brightness Adjustment: Introduced lighting variations to mimic real-world conditions.

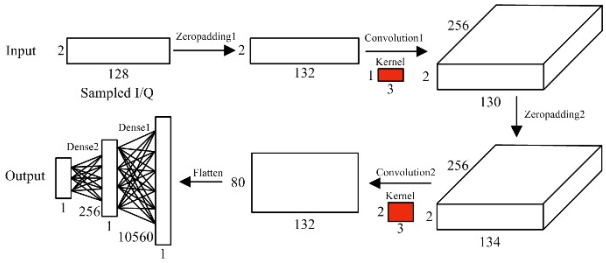
4.3 Deep Learning Architectures

We implemented three deep learning models—CNN, ResNet, and ViT—to evaluate their effectiveness in crack detection.

4.3.1 Convolutional Neural Network (CNN)

CNNs are widely used in image classification due to their ability to capture spatial features through convolutional layers. Our CNN architecture consists of:

- Three convolutional layers (filters: 32, 64, 128), each followed by ReLU activation and MaxPooling.
- Flatten layer to convert feature maps into a 1D vector.
- Fully connected layers leading to a sigmoid activation for binary classification.

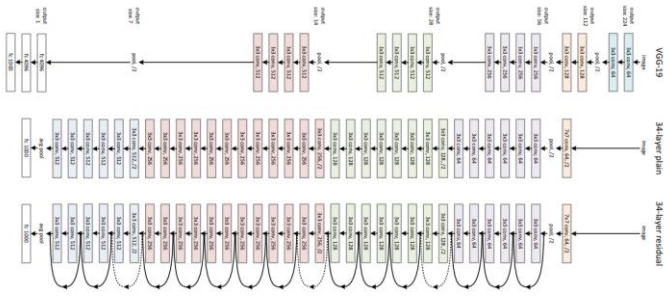


CNN layer architecture [12]

4.3.2 ResNet

Residual Networks (ResNets) are a class of deep neural networks that address the vanishing gradient problem, enabling the training of substantially deeper architectures. For binary image classification tasks, a typical ResNet architecture can be adapted as follows:

- Input Layer: Accepts images with a specified resolution, e.g., 224x224 pixels, and three-color channels (RGB).
- Initial Convolutional Layer**:** Performs convolution with a 7x7 kernel and 64 filters, followed by a stride of 2, and includes batch normalization and ReLU activation.
- Residual Blocks**:** A sequence of residual blocks, each containing convolutional layers with batch normalization and ReLU activation. Shortcut connections bypass these layers to facilitate gradient flow. The number of blocks can vary depending on the desired depth of the network.
- Global Average Pooling Layer: Reduces each feature map to a single value, resulting in a 1D vector.
- Fully Connected Layer: A dense layer with a single neuron and a sigmoid activation function to output probabilities for the two classes.



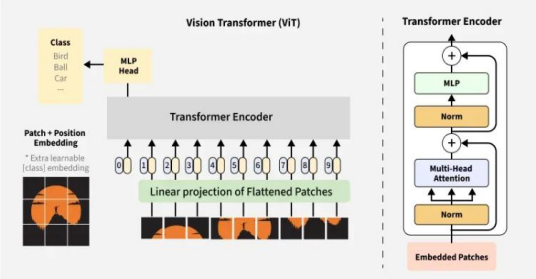
ResNet architecture [13]

4.3.3 Vision Transformer (ViT)

Vision Transformers (ViT) represent a significant departure from conventional convolutional networks by leveraging self-attention mechanisms to capture spatial relationships in images. Unlike CNNs, which process images using localized feature extraction via convolutional filters, ViTs split images into fixed-size patches and process them similarly to words in a sentence in NLP models.

ViT-based architecture for crack detection consists of:

- Patch Embedding Layer:
  The input image (224×224 pixels) is divided into fixed-size patches (16×16). Each patch is flattened and passed through a linear projection layer to create patch embeddings.
- Positional Encoding:
  Since Transformers lack built-in spatial awareness, positional embeddings are added to retain spatial information.
- Transformer Encoder Blocks:
  A series of multi-head self-attention layers process the patch embeddings.
  Feed-forward neural networks (MLPs) follow each attention block.
- **Fully Connected Layer:**
  A dense layer with sigmoid activation predicts the probability of a crack being present (binary classification).



ViT architecture [14]

6. EXPERIMENTAL RESULTS

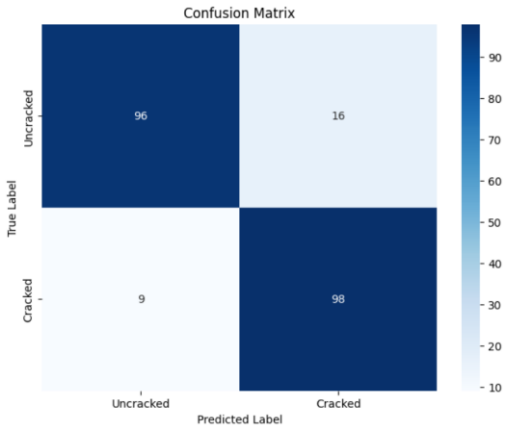
The models were trained for 20 epochs, with early stopping enabled to prevent overfitting. Below are the training and validation accuracy/loss curves to visualize the learning process.

6.2 Confusion Matrix Analysis

After training, the models were evaluated on the test uncracked. Below are the confusion metrics:

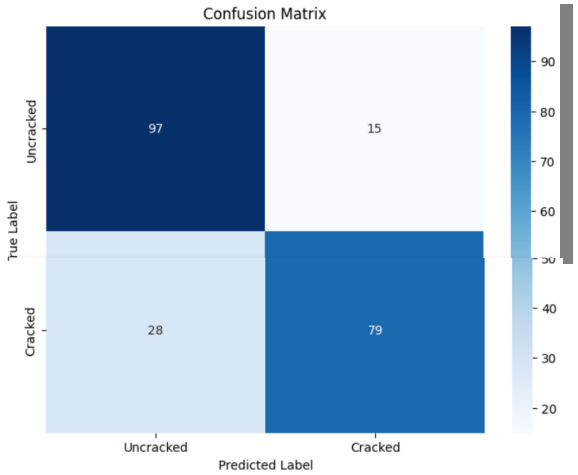
CNN:

accuracy	0.89	219		
macro avg	0.89	0.89	0.89	219
weighted avg	0.89	0.89	0.89	219



ResNet:

accuracy	0.80	219		
macro avg	0.81	0.80	0.80	219
weighted avg	0.81	0.80	0.80	219



ViT:
Accuracy: 0.8904
Precision: 0.9882
Recall: 0.7850
F1 Score: 0.8750

6.3 Models Summary

Model	Accuracy	Precision	Recall	F1-score
CNN	89.00%	89.00%	89.00%	89.00%
ResNet50	80.00%	81.00%	80.00%	80.00%
ViT	89.04%	98.82%	78.50%	87.50%

6

4

6.4 Key Observation

1- CNN Performance

- Achieved an accuracy of 89.0%, with balanced precision, recall, and F1-score (all at 89%).
- The confusion matrix indicates some false positives and false negatives, meaning that CNN struggles slightly in distinguishing cracked from uncracked pipes.
- It provides stable performance but lacks the advanced feature extraction capabilities of deeper architectures like ResNet and ViT.

2- ResNet50 Performance

- Surprisingly, ResNet50 performed the worst among the three models, with an accuracy of 80.0%.
- The confusion matrix shows a high number of misclassified cracked pipes, leading to a lower recall (80%) compared to CNN and ViT.
- While ResNet50 typically excels in image classification tasks, its relatively poor performance in this case may indicate overfitting, suboptimal hyperparameters, or insufficient fine-tuning for the dataset.

3- ViT Performance

- ViT achieved an accuracy of 89.04%, slightly higher than CNN and significantly better than ResNet50.
- It had the highest precision (98.82%), meaning that when ViT predicts a crack, it is very confident in its classification.
- However, recall (78.50%) was lower, indicating that ViT misses some cracked samples, possibly due to its reliance on global attention mechanisms rather than localized feature extraction like CNNs.

6.5 Overall Findings

- CNN and ViT performed similarly in terms of accuracy, but ViT demonstrated higher precision and lower recall.
- ResNet50 did not perform as expected, likely due to training inefficiencies or dataset-specific challenges.
- The results suggest that ViT is excellent at correctly identifying cracked pipes (high precision), while CNN balances recall and precision better.
- For industrial applications, choosing between ViT and CNN depends on whether minimizing false positives (ViT) or capturing all cracked pipes (CNN) is the priority.

7. CONCLUSION

This research explored the application of deep learning models—CNN, ResNet50, and Vision Transformer (ViT)—for automatic crack detection in pipelines using a dataset of 2,000 images (1,000 cracked, 1,000 uncracked). The study aimed to determine the most effective model by evaluating performance based on accuracy, precision, recall, F1-score, and confusion matrices.

The results of this study suggest that deep learning can significantly improve pipeline crack detection, reducing reliance on manual inspection and traditional non-destructive testing (NDT) methods. Implementing automated crack detection systems using ResNet50 can lead to:

1. Faster and more accurate defect detection, reducing downtime and maintenance costs.
2. Improved safety and risk management by detecting cracks before they lead to catastrophic failures.
3. Integration with real-time monitoring systems for automated pipeline inspections in industrial environments.

ACKNOWLEDGMENTS

The author gratefully acknowledges the support provided by the Undergraduate Research Office (URO) through the Guided Research program. We also extend our appreciation to the IRC for Sustainable Energy & Power Systems at KFUPM for their essential contributions during the research period. Their provision of resources and research facilities was critical to the successful execution of this study.

REFERENCES

[1] M. Islam, "Oil pipeline," *ScienceDirect*, 2023. [Online]. Available: https://www.sciencedirect.com/topics/engineering/oil-pipeline.

[2] S. Vishnuvardhan, A. Ramachandra, and A. Choudhary, "Optimizing pipeline integrity management: A multi-criteria decision analysis approach," *ScienceDirect*, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0308016122002381.

[3] R. Ali and M. Shoaib, "Crack detection," *ScienceDirect*, 2022. [Online]. Available: https://www.sciencedirect.com/topics/engineering/crack-detection.

[4] IEEE Xplore, "Deep learning-based crack detection for structural health monitoring," *IEEE Xplore*, 2024. https://ieeexplore.ieee.org/abstract/document/10589380.

[5] ScienceDirect**,** ImageandVisionComputing**,** *ScienceDirect*, 2022. [Online]. Available: https://www.sciencedirect.com/topics/computer-science/residual-neural-network#:~:text=A%20Residual%20Neural%20Network%20(ResNet,of%20information%20from%20input%20data.

[6] M. Touvron, H. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," *arXiv preprint arXiv:2010.11929*, 2020. [Online]. Available: https://arxiv.org/pdf/2010.11929.

[7] U.S. Department of Energy, "MFL Inspection and its limitations," *National Energy Technology Laboratory (NETL)*, 2018. [Online]. Available:

https://netl.doe.gov/sites/default/files/2018-03/MFL_inspection_r4.pdf?utm_source.
[8] A. K. Dubey and M. Kumar, "Magnetic Barkhausen noise technique for microstructural and stress state evaluation," *Applied Sciences*, vol. 10, no. 24, p. 8938, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/24/8938.
[9] F. Ahmad, S. Iqbal, and M. Hussain, "Infrared thermography for subsurface defect detection: A review," *Sensors*, vol. 22, no. 18, p. 7098, 2022. https://www.mdpi.com/1424-8220/22/18/7098.
[10] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," *Scientific Research Publishing*, 2023. [Online]. Available: https://www.scirp.org/journal/paperinformation?paperid=136828&.
[11] DigitalOcean, "Popular deep learning architectures: AlexNet, VGG, and GoogLeNet," *DigitalOcean Tutorials*, 2023. [Online]. Available: https://www.digitalocean.com/community/tutorials/popular-deep-learning-architectures-alexnet-vgg-googlenet.[Online]. Available:
[12] ResearchGate, "The main structure of CNN for VHF signal modulation classification," *ResearchGate*, 2019.https://www.researchgate.net/figure/The-main-structure-of-CNN-for-VHF-signal-modulation-classification_fig2_329479740.
[13] GeeksforGeeks, "Residual Networks (ResNet) in deep learning," *GeeksforGeeks*, 2023. [Online]. Available: https://www.geeksforgeeks.org/residual-networks-resnet-deep-learning/.
[14] GeeksforGeeks, "Vision Transformer (ViT) architecture," *GeeksforGeeks*, 2023. [Online]. Available: https://www.geeksforgeeks.org/vision-transformer-vit-architecture/.