Proceedings Paper:
George, A., Shepherd, W., Tait, S. et al. (2 more authors) (2025) A deep learning benchmark analysis of the publicly available WRc dataset for sewer defect classification. In: Proceedings of 21st Computing & Control for the Water Industry Conference (CCWI 2025). CCWI 2025 - 21st Computing & Control for the Water Industry Conference, 01-03 Sep 2025, Sheffield, UK. University of Sheffield.

https://doi.org/10.15131/shef.data.29920874.v1

A Deep Learning Benchmark Analysis of the Publicly Available WRc Dataset for Sewer Defect Classification

Alex George[1], Will Shepherd[2], Simon Tait[2], Lyudmila Mihaylova[1], Sean Anderson[1]
[1] School of Electrical and Electronic Engineering, University of Sheffield
[2] School of Mechanical, Aerospace and Civil Engineering, University of Sheffield
[1]ageorge4@sheffield.ac.uk

ABSTRACT

Deep learning has the potential to transform sewer pipe inspection by automating the process, which could improve efficiency and consistency. However, progress has been hampered by limited publicly available, well-annotated benchmark datasets for defect classification. To address this gap, we present a comprehensive analysis using the publicly available Water Research Centre (WRc) sewer image dataset. We evaluated several deep learning architectures (MobileNet-v2, Inception-ResNet-v2 and ResNet-18) across key performance metrics such as accuracy and F1-score, with Top-1 accuracies ranging from 61.54% to 71.61% and Top-3 accuracies ranging from 86.88% to 92.61%. This research contributes to a reproducible performance baseline, enabling rigorous comparison of different models and serves as a foundation for future research in developing AI-assisted inspection systems.

Keywords: CCTV Inspection, Deep Learning, Sewer Defect Detection

INTRODUCTION

Routine sewer infrastructure maintenance is vital to ensure public health and avoid costly repairs. One critical task is the inspection of sewer pipes, traditionally performed through manual review of closed-circuit television (CCTV) footage. This process, however, is time-consuming and prone to human error due to fatigue, lighting variability, and complex defect appearances [1]. The water industry is therefore exploring automated methods for assessing pipe conditions to overcome these challenges [2]. Deep learning algorithms have attracted considerable interest for automated sewer inspection because of their capability to learn complex features from raw images, enabling them to process large and diverse datasets [3,4]. Recent studies have demonstrated the effectiveness of deep learning in sewer condition assessment, achieving good accuracy in defect classification from CCTV imagery [5,6].

Many image-based sewer defect classification systems use datasets with varying classes and class distributions. However, the lack of standardised datasets and evaluation protocols impedes progress in this field [7]. Recent efforts by Haurum and Moeslund [8] and the Water Research Centre (WRc) have begun to address this, with WRc releasing a dedicated publicly available sewer defect dataset for training machine learning systems [9].

A public benchmark analysis of the WRc dataset for defect classification is not yet available. To address this gap, a deep learning analysis has been applied in this paper to the WRc dataset. Our initial results provide insights into the feasibility of automated defect classification using this dataset and highlight areas for future research and development.

THE WRc DATASET

WRc compiled the dataset from previously coded CCTV survey footage contributed by seven UK water utilities: United Utilities, Thames Water, South West Water, Dŵr Cymru, Scottish Water, Severn Trent Water and Yorkshire Water. WRc used 15 water engineers and technical consultants from its technical consulting and catchment modelling departments to process the CCTV images. Their task was to complete the defect classification process and select optimal images for the training library.

At present, the WRc dataset consists of 27,257 images under 72 unique Manual of Sewer Condition Classification (MSCC) defect codes [10], with varying numbers of samples per defect, from over 1000 (e.g. Crack Longitudinal (CL)) to as few as single digits (e.g. Exfiltration (EX)). Each image is also accompanied by metadata detailing attributes such as the sewer pipe material, pipe diameter and the approximate location of the defect within the frame. MSCC is the UK CCTV defect coding system, but can be converted to the European standard codes, and is similar to many other coding methods around the globe, as discussed in [11]. The dataset can be accessed through the Spring platform [12], by registering and selecting the AI and Sewer case study.

BENCHMARK ANALYSIS METHODS

DATASET PREPARATION

A subset of the WRc dataset was created to ensure sufficient representation from each defect class. The 72 MSCC defect codes were first grouped into 31 primary classes by combining related defect codes based on their primary category, e.g. combining Crack Circumferential (CC) with Crack Longitudinal (CL) into a single ‘Crack’ class - this resulted in Fig. 1. Then, primary classes with over 800 images were selected for incslusion to ensure sufficient representations per class for training a classifier. This resulted in the inclusion of 13 primary classes and the exclusion of 18 primary classes (Fig. 1). Examples of each included class are shown in Fig. 2. To produce the actual dataset for processing, 1000 images were sampled from each of these primary classes to create a balanced dataset for classification (upsampling the ‘Defective Connection’ and ‘Deformed’ classes that had 839 and 847 images respectively by duplication of randomly selected images). The dataset was randomly split in the ratio 70:10:20 per class for training, validation and testing purposes.

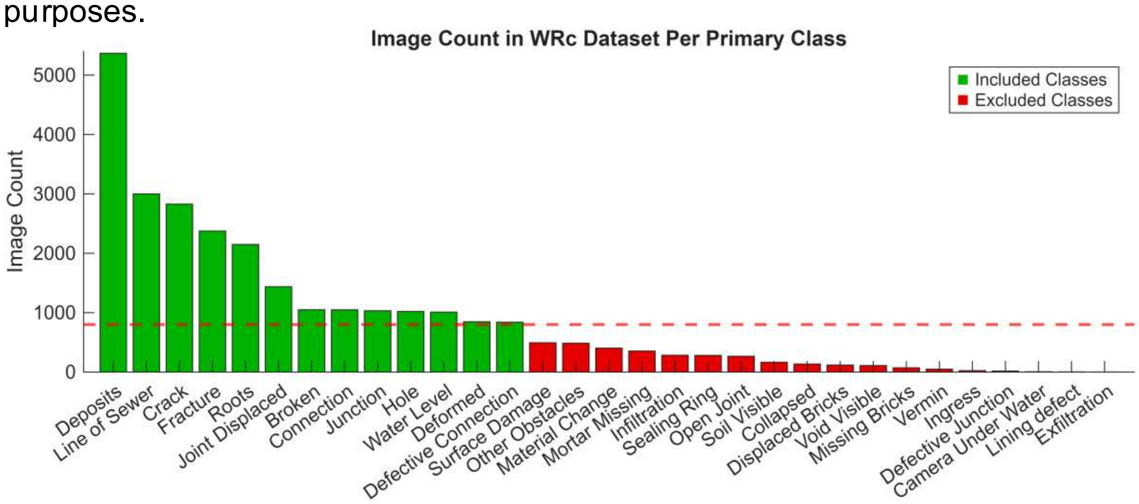


Figure 1. Image count in the WRc dataset per primary class. The dashed red line marks the threshold where classes with at least 800 images are considered.

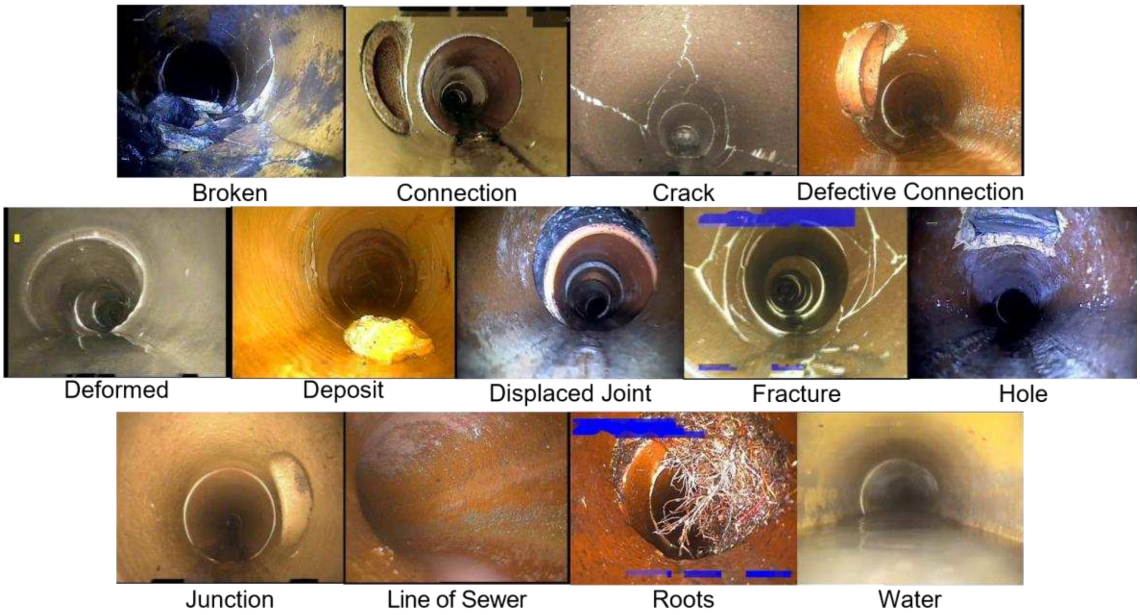


Figure 2. Example images of the 13 classes used for classification in this study.

MODEL SELECTION AND TRAINING

Transfer learning with the WRc data was applied to three standard deep learning models, pre-trained on the ImageNet database [13]: MobileNet-v2 (small network), ResNet-18 (medium-sized network) and Inception-Resnet-v2 (large network). The choice of the three models was based on their demonstrated success in various image classification tasks and their proven ability to generalise well when fine-tuned for specific datasets. A dropout layer was added just before the final fully connected layer to prevent overfitting. Images were also resized to scale to the input size of the pre-trained networks, maintaining the aspect ratio of the original image using zero padding. To ensure a fair comparison, the general training procedure was the same for all three networks. Standard data augmentation techniques were used to improve model generalisation and robustness, such as translation, rotation, flipping, brightness, blurring and noise adjustment, which were applied randomly to 50% of the training dataset. All models were trained for 10 epochs using the Adam optimiser with focal loss, starting from an initial learning rate of 0.0001 with a decay factor of 0.1 for every 5 epochs. Convergence was monitored using the validation dataset. A mini-batch size of 32 was used, along with a dropout probability of 0.3 and an L2 regularisation rate of 0.0001. The models were trained and evaluated using an NVIDIA GeForce 4070 GPU.

HIGHER-LEVEL DATA GROUPING

We investigated grouping visually similar classes that led to frequent confusions, with groups defined as: 1. Crack plus Fracture, 2. Connection plus Defective Connection, and 3. Broken plus Deformed plus Hole. This grouping was done to simplify the classification task, based on expert opinion and a review of defect characteristics. These 3 higher-level groupings were combined with the remaining 6 original classes in Fig. 1, giving 9 classes in total. The effect of grouping the data was analysed in two ways: one via post-processing the model predictions into the defined groups and one via training a ResNet-18 model directly on the grouped data (ResNet-18-Grouped). Each grouped class was randomly sampled to contain 1000 images, split as before in the ratio of 70:10:20 for training, validation and testing.

RESULTS AND DISCUSSION

Among the models evaluated, MobileNet-v2 achieved the lowest accuracy (61.54%), while Inception-ResNet-v2 achieved the highest (66.92%). ResNet-18 achieved the best trade-off in accuracy (64.73%) and prediction time (0.006s to process an image) (Table 1). F1-scores followed a similar profile to accuracy (Table 1). All models showed strong Top-3 accuracy (i.e. when the true class is among the top three predictions), with ResNet-18 obtaining 88.08% (Table 1). The trained models are shared online at GitHub [14].

Certain classes showed low per-class accuracies due to visually similar characteristics that caused confusion, such as Crack vs. Fracture (Fig. 3(a)), Connection vs. Defective Connection (Fig. 3(b)), and Broken, Hole, and Deformed (Fig. 3(c)) - also see the ResNet-18 confusion matrix in Fig. 4. Some images contained more than one feature, making multi-class classification challenging as the model attempted to predict a single best category. Therefore, confidence was spread across relevant classes. For example, Fig. 3(d) shows an image labelled Displaced Joint, but the presence of structures resembling roots raised the Roots class confidence. In addition, some noisy images were difficult to classify reliably, e.g., Fig. 3(e), where the model failed to identify the defect as Roots. These examples demonstrate some of the challenges in working with this dataset in sewer image classification.

Grouping visually similar classes improved the accuracy, reducing confusion (Table 1 and Fig. 5). We grouped classes in two ways: one via post-processing the baseline ResNet-18 predictions, which achieved an accuracy of 73.77%, and the other by directly training on grouped classes (ResNet-18-Grouped), which led to a slightly lower accuracy of 71.61% (see Figs. 5(a) and 5(b)).

Table 1. Classification results on the test dataset

Architecture	Number of params.	Pred. time (s)	Acc. (%)	Top-3 Acc. (%)	Grouped Acc. (%)	Macro Prec. (%)	Macro Recall (%)	Macro F1-score (%)
MobileNet-v2	2.21M	0.018	61.54	86.88	72.00	61.99	61.54	61.70
ResNet-18	11.17M	0.006	64.73	88.08	73.77	64.90	64.73	64.69
Inception-ResNet-v2	54.28M	0.097	66.92	90.42	76.88	67.25	66.92	67.02
ResNet-18 - Grouped	11.17M	0.006	71.61	92.61	71.61	71.54	71.61	71.48

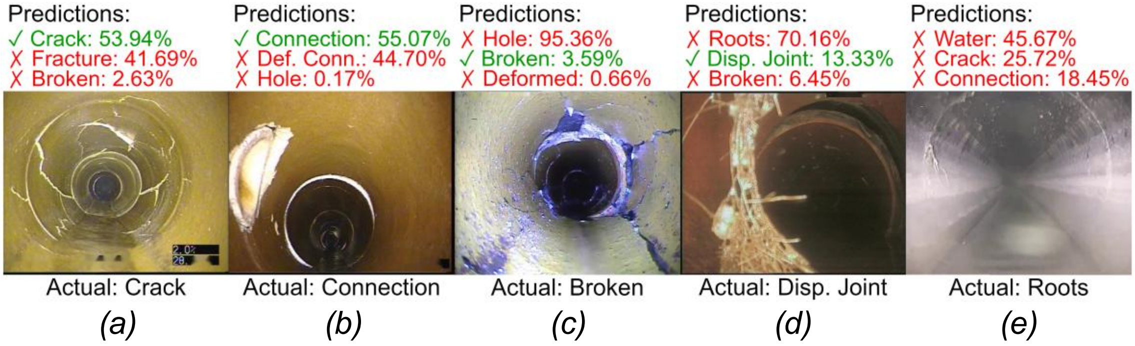
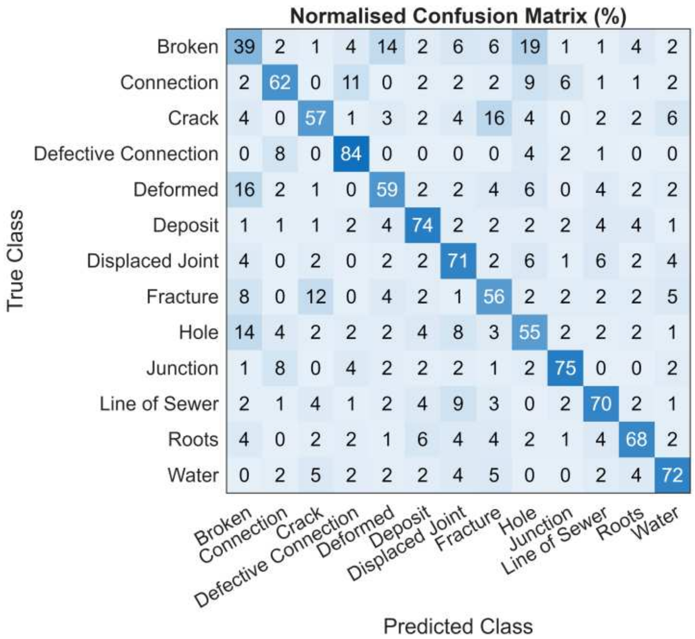


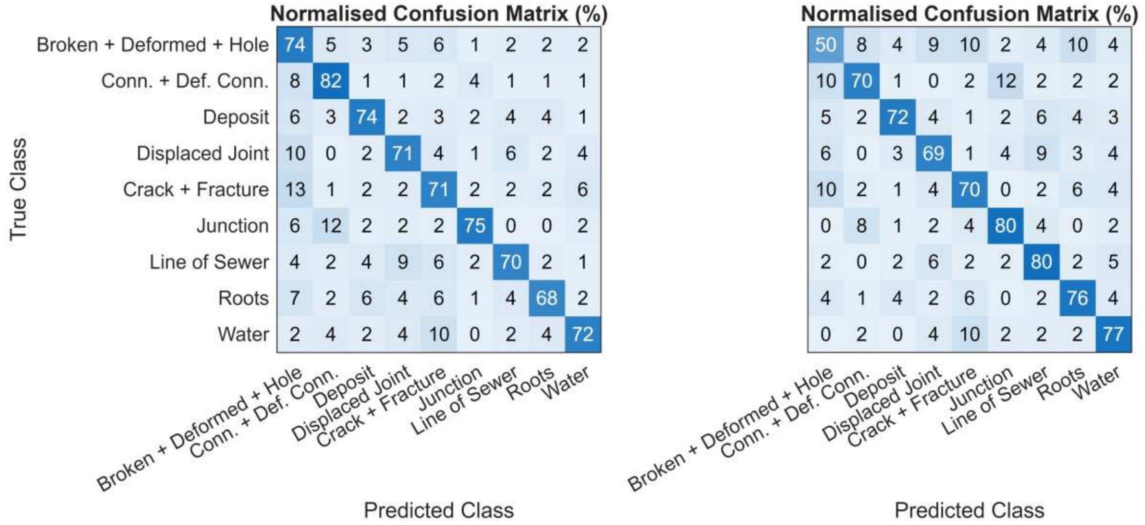
Figure 3. Examples of Top-3 classification predictions using ResNet-18.



Normalised Confusion Matrix (%)

True Class \ Predicted Class	Broken	Connection	Crack	Defective Connection	Deformed	Deposit	Displaced Joint	Fracture	Hole	Junction	Line of Sewer	Roots	Water
Broken	39	2	1	4	14	2	6	6	19	1	1	4	2
Connection	2	62	0	11	0	2	2	2	9	6	1	1	2
Crack	4	0	57	1	3	2	4	16	4	0	2	2	6
Defective Connection	0	8	0	84	0	0	0	0	4	2	1	0	0
Deformed	16	2	1	0	59	2	2	4	6	0	4	2	2
Deposit	1	1	1	2	4	74	2	2	2	2	4	4	1
Displaced Joint	4	0	2	0	2	2	71	2	6	1	6	2	4
Fracture	8	0	12	0	4	2	1	56	2	2	2	2	5
Hole	14	4	2	2	2	4	8	3	55	2	2	2	1
Junction	1	8	0	4	2	2	2	1	2	75	0	0	2
Line of Sewer	2	1	4	1	2	4	9	3	0	2	70	2	1
Roots	4	0	2	2	1	6	4	4	2	1	4	68	2
Water	0	2	5	2	2	2	4	5	0	0	2	4	72

Figure 4. Confusion matrix using ResNet-18 on the test dataset.



Normalised Confusion Matrix (%)

True Class \ Predicted Class	Broken + Deformed + Hole	Conn. + Def. Conn.	Deposit	Displaced Joint	Crack + Fracture	Junction	Line of Sewer	Roots	Water
Broken + Deformed + Hole	74	5	3	5	6	1	2	2	2
Conn. + Def. Conn.	8	82	1	1	2	4	1	1	1
Deposit	6	3	74	2	3	2	4	4	1
Displaced Joint	10	0	2	71	4	1	6	2	4
Crack + Fracture	13	1	2	2	71	2	2	2	6
Junction	6	12	2	2	2	75	0	0	2
Line of Sewer	4	2	4	9	6	2	70	2	1
Roots	7	2	6	4	6	1	4	68	2
Water	2	4	2	4	10	0	2	4	72

(a)

Normalised Confusion Matrix (%)

True Class \ Predicted Class	Broken + Deformed + Hole	Conn. + Def. Conn.	Deposit	Displaced Joint	Crack + Fracture	Junction	Line of Sewer	Roots	Water
Broken + Deformed + Hole	50	8	4	9	10	2	4	10	4
Conn. + Def. Conn.	10	70	1	0	2	12	2	2	2
Deposit	5	2	72	4	1	2	6	4	3
Displaced Joint	6	0	3	69	1	4	9	3	4
Crack + Fracture	10	2	1	4	70	0	2	6	4
Junction	0	8	1	2	4	80	4	0	2
Line of Sewer	2	0	2	6	2	2	80	2	5
Roots	4	1	4	2	6	0	2	76	4
Water	0	2	0	4	10	2	2	2	77

(b)

Figure 5. Confusion matrices for grouped classes on the test dataset: (a) Grouping applied to the baseline ResNet-18 predictions in post-processing; (b) ResNet-18 trained on grouped labels (ResNet-18-Grouped).

CONCLUSIONS

This benchmark analysis has characterised the effectiveness of the WRc sewer defect dataset for training deep learning systems for automated sewer inspection. Among the models tested, ResNet-18 provided a good balance of accuracy and speed, achieving 64.73% accuracy with a fast prediction time of just 0.006 seconds. Additionally, all models showed strong Top-3 accuracy (92.61% with ResNet-18-Grouped), indicating that even when the top prediction is incorrect, the true class is often among the top three, which is promising for real-world applications where a shortlist of probable defects is sufficient. Grouping the classes with similar characteristics further increased the accuracy to 73.77% on ResNet-18 and 71.61% on ResNet-18-Grouped, thereby removing the need to distinguish between categories with frequent confusion. The proposed approach with ensemble properties has the potential for automated analysis of pipe networks. Future work will focus on detecting multiple defects in single-label images, especially with low resolution, noise and obscure fine details.

ACKNOWLEDGEMENTS

This work is supported by the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement No. 101189847 - Pipeon.

REFERENCES

[1] J. Dirksen et al., "The consistency of visual sewer inspection data," *Struct. Infrastruct. Eng.*, vol. 9, no. 3, pp. 214–228, Mar. 2013.
[2] R. Rayhana, Y. Jiao, A. Zaji, and Z. Liu, "Automated vision systems for condition assessment of sewer and water pipelines," *IEEE Trans. Automat. Sci. Eng.*, vol. 18, no. 4, pp. 1861–1878, Oct. 2021.
[3] Y. Li, H. Wang, L. M. Dang, H. K. Song, and H. Moon, "Vision-based defect inspection and condition assessment for sewer pipes: A comprehensive survey," *Sensors*, vol. 22, no. 7, Art. no. 2722, Apr. 2022.
[4] S. Pouyanfar et al., "A survey on deep learning: Algorithms, techniques, and applications," *ACM Comput. Surv.*, vol. 51, no. 5, Art. no. 92, Sep. 2018.
[5] D. Meijer, L. Scholten, F. Clemens, and A. Knobbe, "A defect classification methodology for sewer image sets with convolutional neural networks," *Autom. Constr.*, vol. 104, pp. 281–298, Aug. 2019.
[6] D. Li, A. Cong, and S. Guo, "Sewer damage detection from imbalanced CCTV inspection data using deep convolutional neural networks with hierarchical classification," *Autom. Constr.*, vol. 101, pp. 199–208, May 2019.
[7] J. B. Haurum and T. B. Moeslund, "A survey on image-based automation of CCTV and SSET sewer inspections," *Autom. Constr.*, vol. 111, Art. no. 103061, Mar. 2020.
[8] J. B. Haurum and T. B. Moeslund, "Sewer-ML: A multi-label sewer defect classification dataset and benchmark," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13451–13462.
[9] P. Henley, "A global first in sewer inspection launches," wrcgroup.com, Accessed: Feb. 21, 2025. [Online]. Available: https://www.wrcgroup.com/headlines/bite-sized-views/a-global-first-in-sewer-inspection-launches/
[10] Water Research Centre, *Manual of Sewer Condition Classification*, 5th ed., WRc plc, Swindon, U.K., 2013.
[11] F. Boogaard et al., "Investigate the condition of an asset," in *Asset Management of Urban Drainage Systems: If Anything Exciting Happens, We've Done It Wrong!*, F. Cherqui, F. Clemens-Meyer, F. Tscheikner-Gratl, and B. van Duin, Eds., London, U.K.: IWA, 2024, pp. 93–130.
[12] Spring. "Accelerating Water Sector Transformation". Spring-Innovation.co.uk. Accessed: 7 July 2025. [Online.] Available: https://spring-innovation.co.uk/
[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[14] A. George, "WRc-Dataset-Classification", github.com. Accessed: July 10, 2025. [Online]. Available: https://github.com/alexgeorge13/WRc-Dataset-Classification