Using Explanations to Identify Problems and Limitations in AI Models used for Intelligent Maintenance

Giovanni FLOREALE[1], Piero BARALDI[1], Enrico ZIO[2,1], Olga FINK[3],

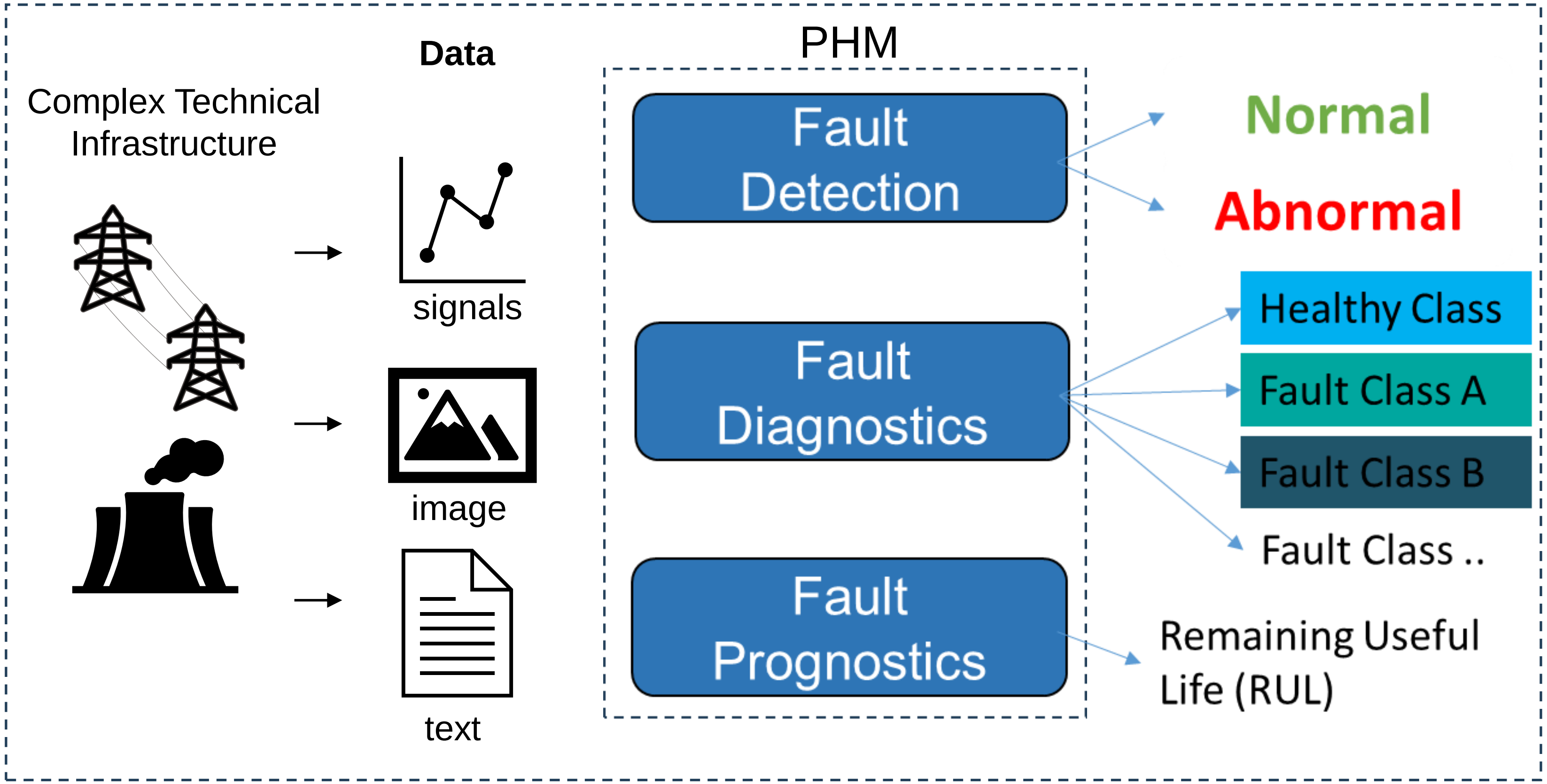
1 Department of Energy, Politecnico di Milano, Milan, Italy

2 MINES Paris-PSL University, Centre de Recherche sur les Risques et les Crises (CRC), Sophia Antipolis, France

3 IMOS, EPFL, Lausanne, Switzerland

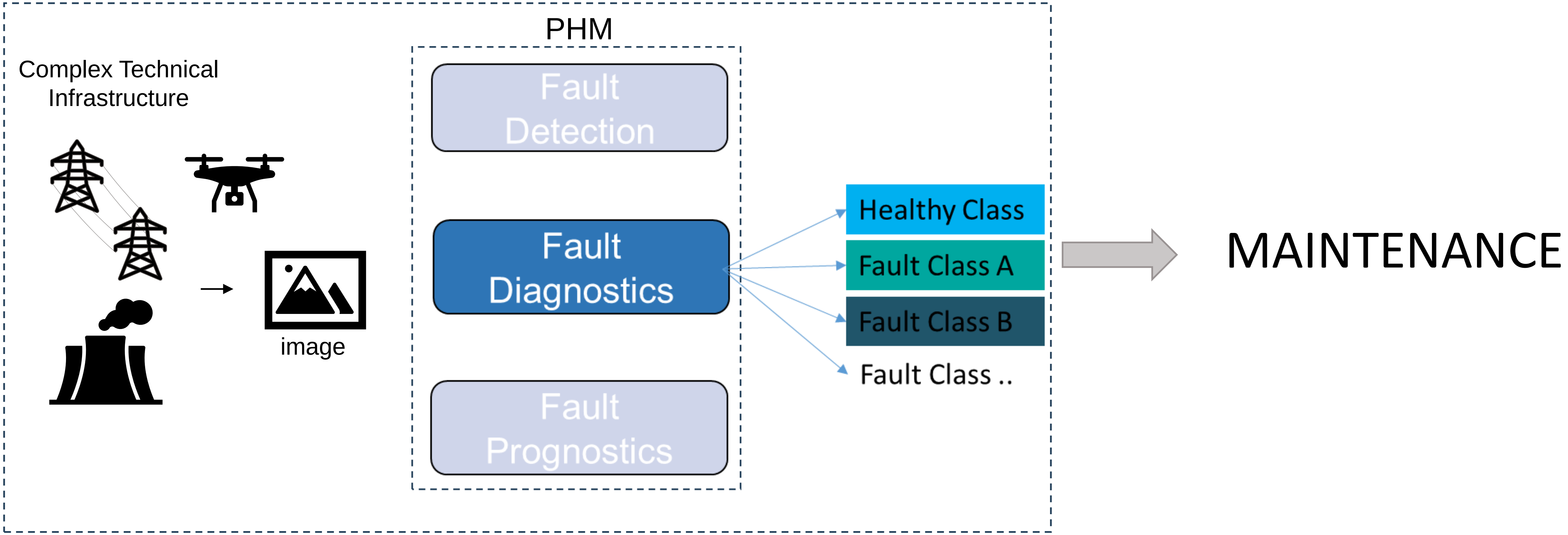
Lausanne, September 12th, 2023

Context: What is the Problem?



→ MAINTENANCE

Context: What is the Problem?



PHM

Complex Technical Infrastructure

image

Fault Detection

Fault Diagnostics

Fault Prognostics

Healthy Class

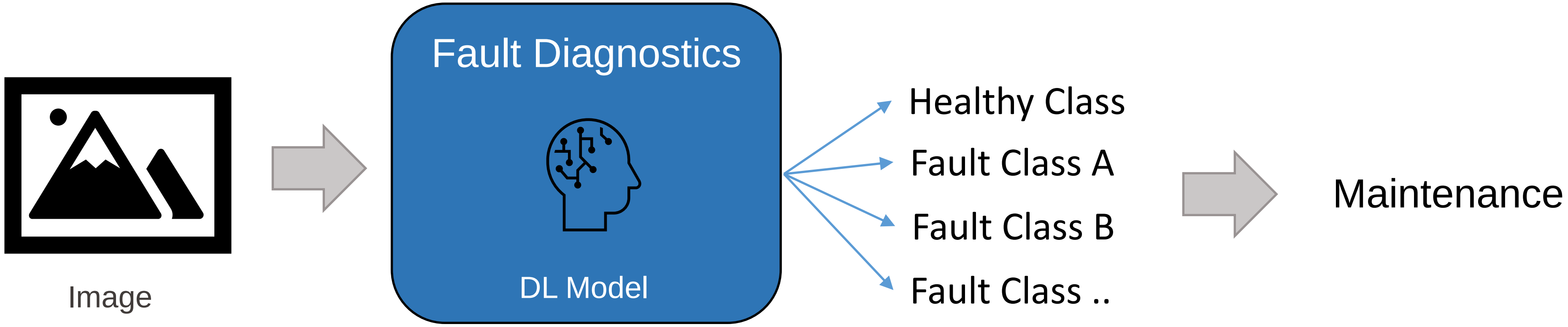
Fault Class A

Fault Class B

Fault Class ..

MAINTENANCE

How is Fault Diagnostics from Images done?



Image

Fault Diagnostics

DL Model

Healthy Class

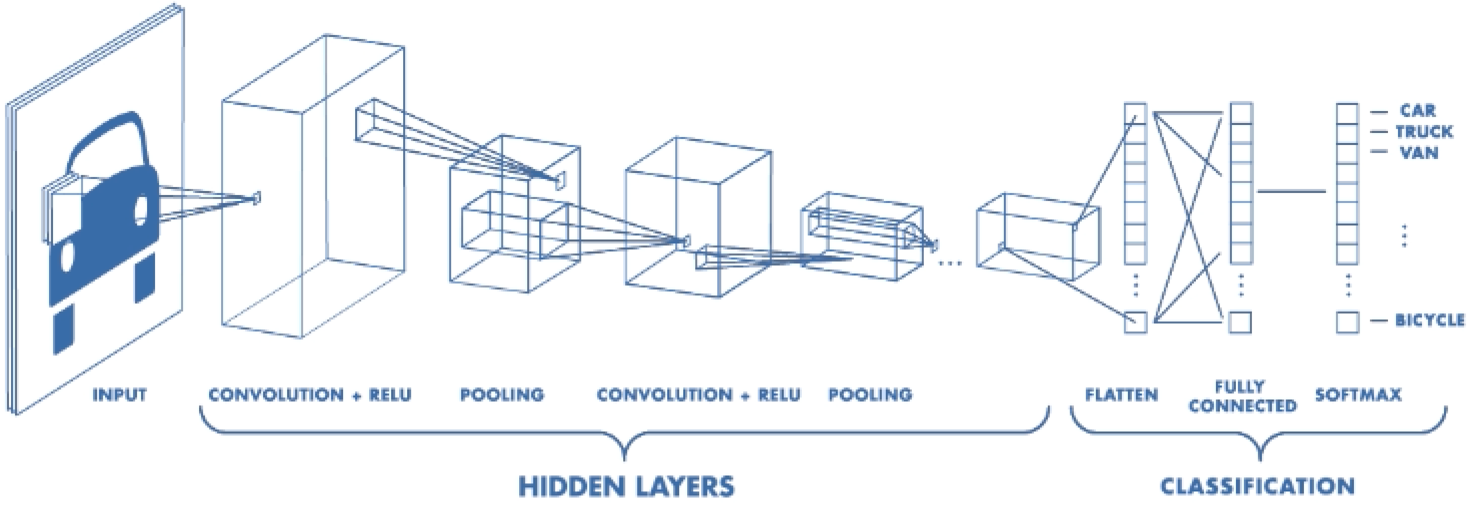
Fault Class A

Fault Class B

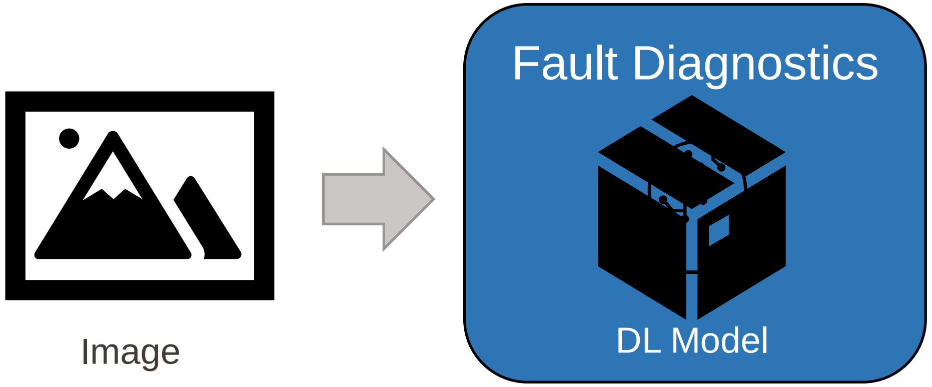
Fault Class ..

Maintenance

e.g. Convolutional Neural Networks

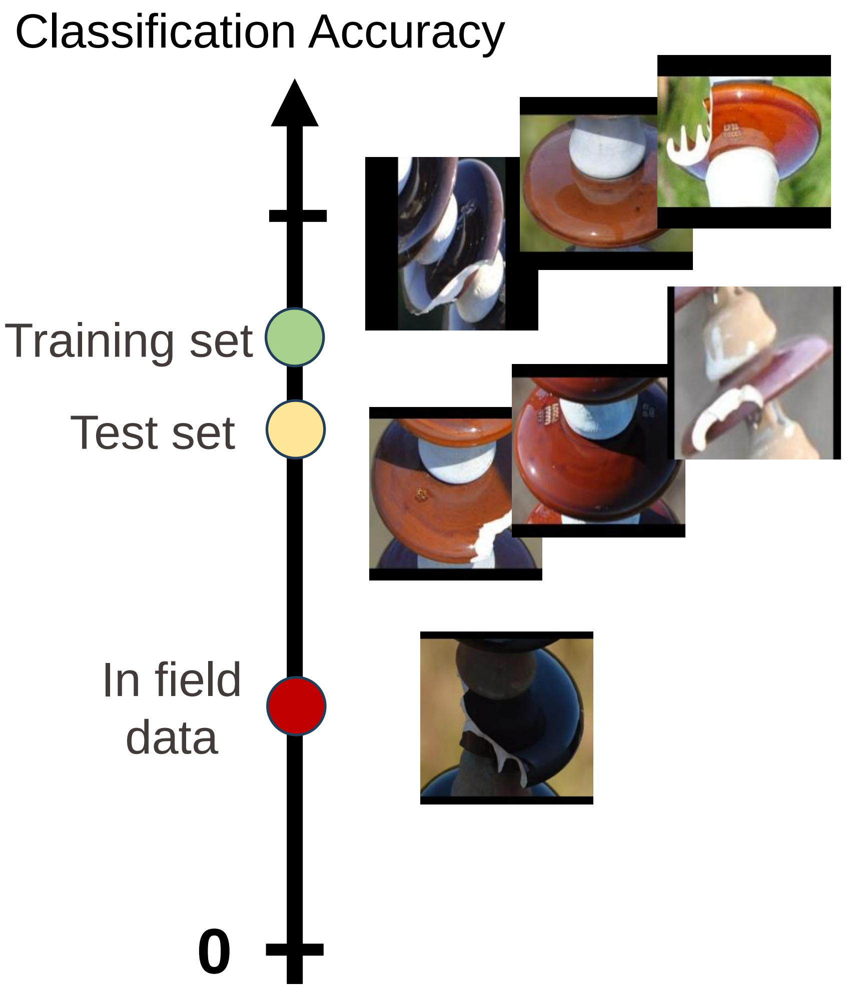


What are the Technical and Scientific Challenges?

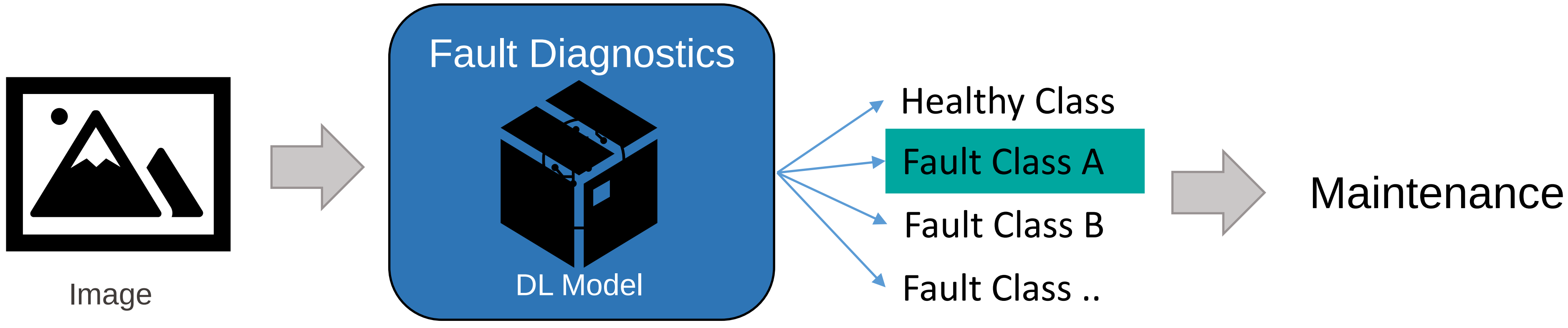


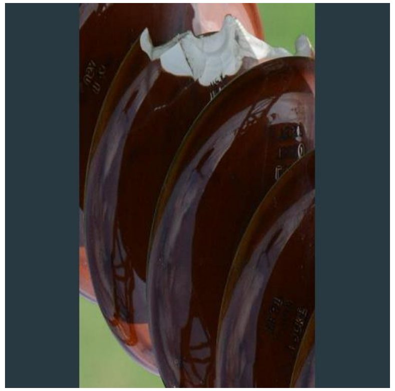
Challenges:

- Black Box
- Performance



Opportunities: What Can Be Done?

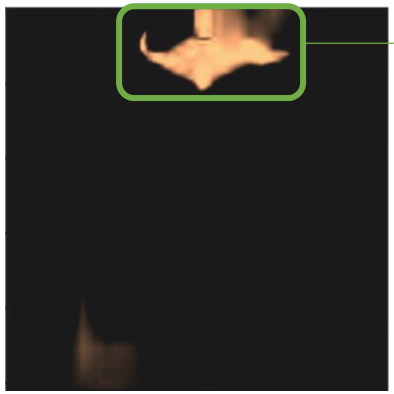




Original image

Model to generate explanations

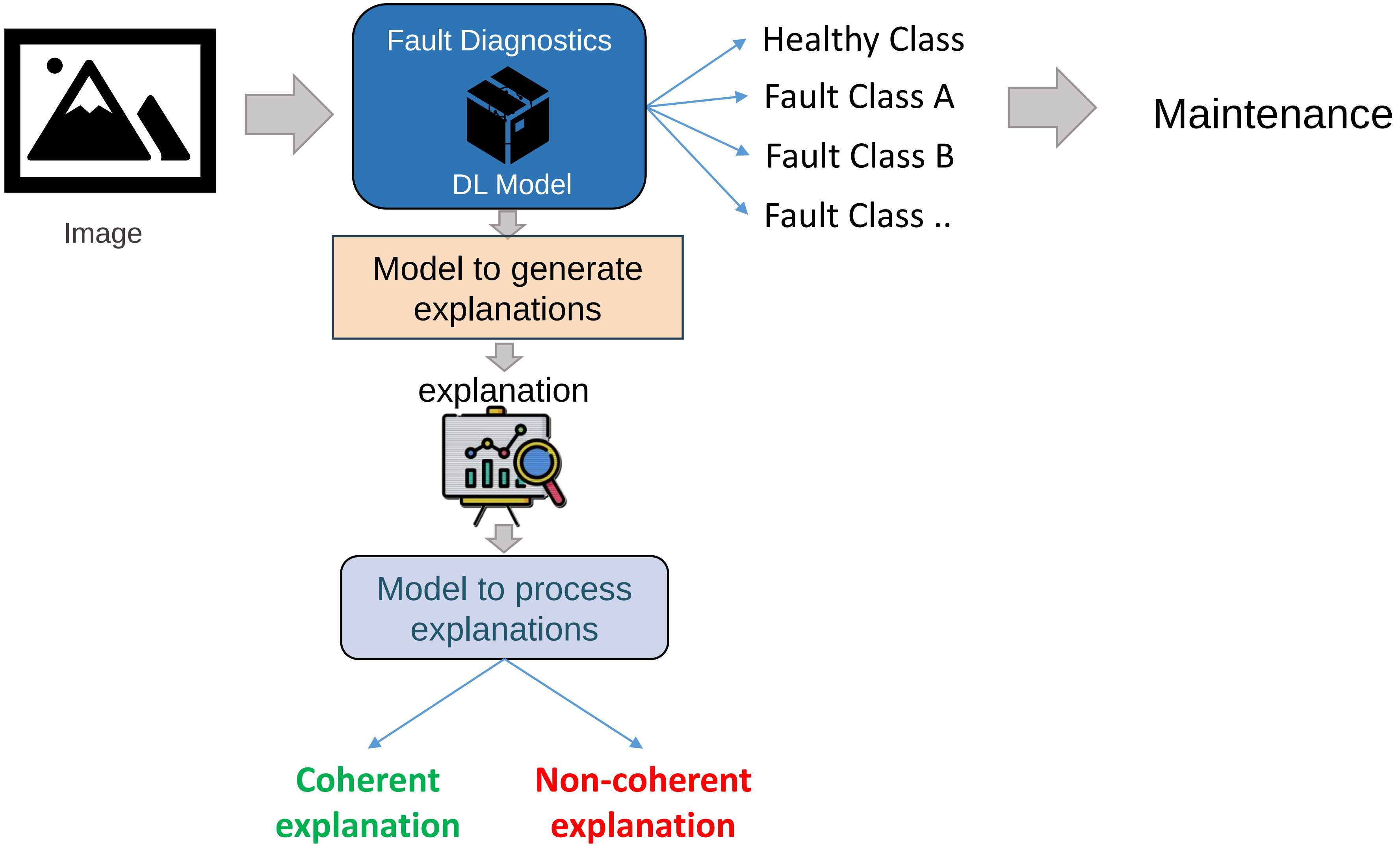
Explanation



Layer-wise Relevance Propagation (LRP)
SHAP
Explainer
Rate Distortion Explanations
CartoonX
…

Most relevant part of the input for the DL model decision on the classification outcome

Objective: What are we Trying to Do?



Image

Fault Diagnostics

DL Model

Healthy Class

Fault Class A

Fault Class B

Fault Class ..

Maintenance

Model to generate explanations

explanation

Model to process explanations

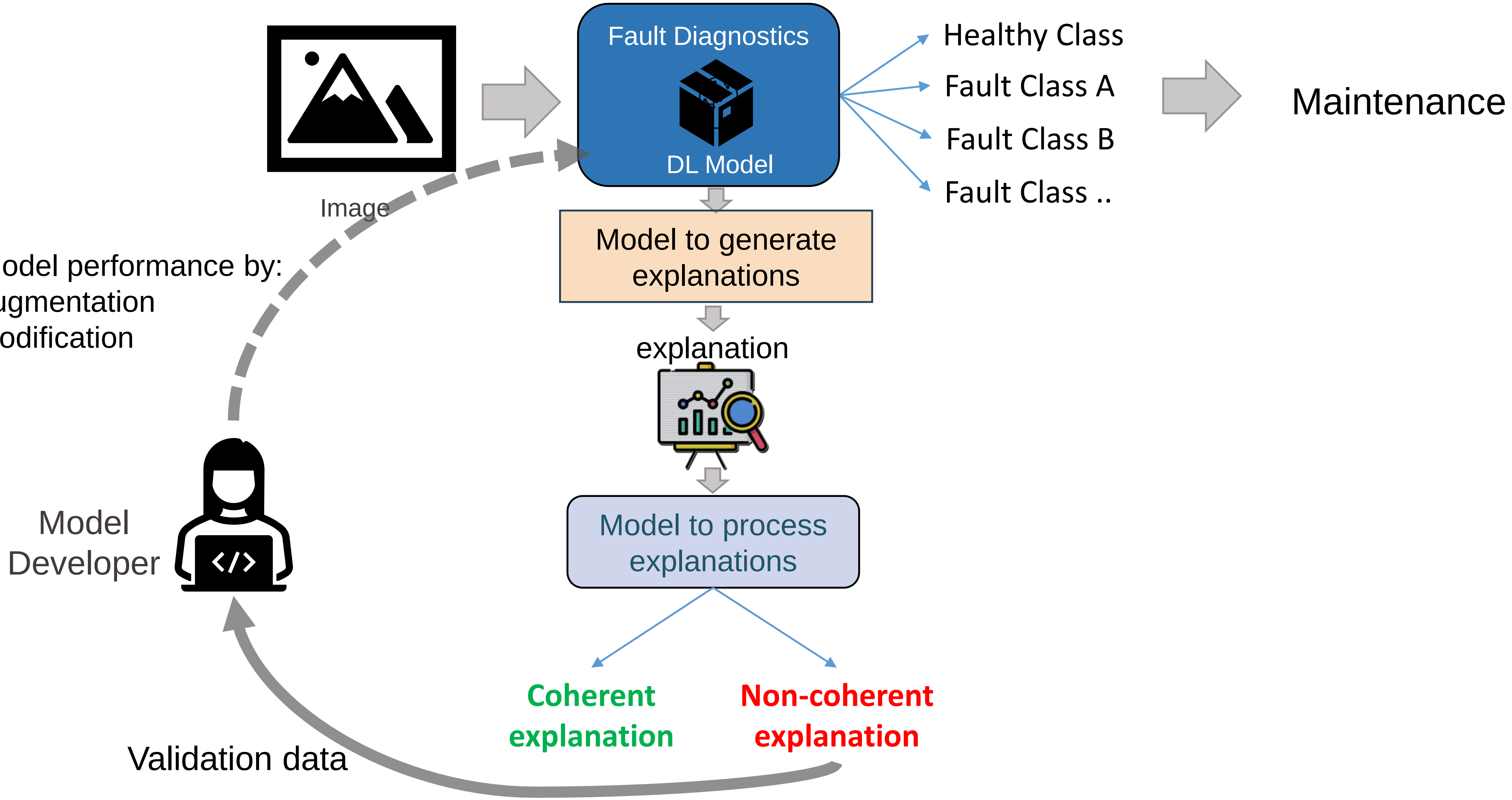
Coherent explanation

Non-coherent explanation

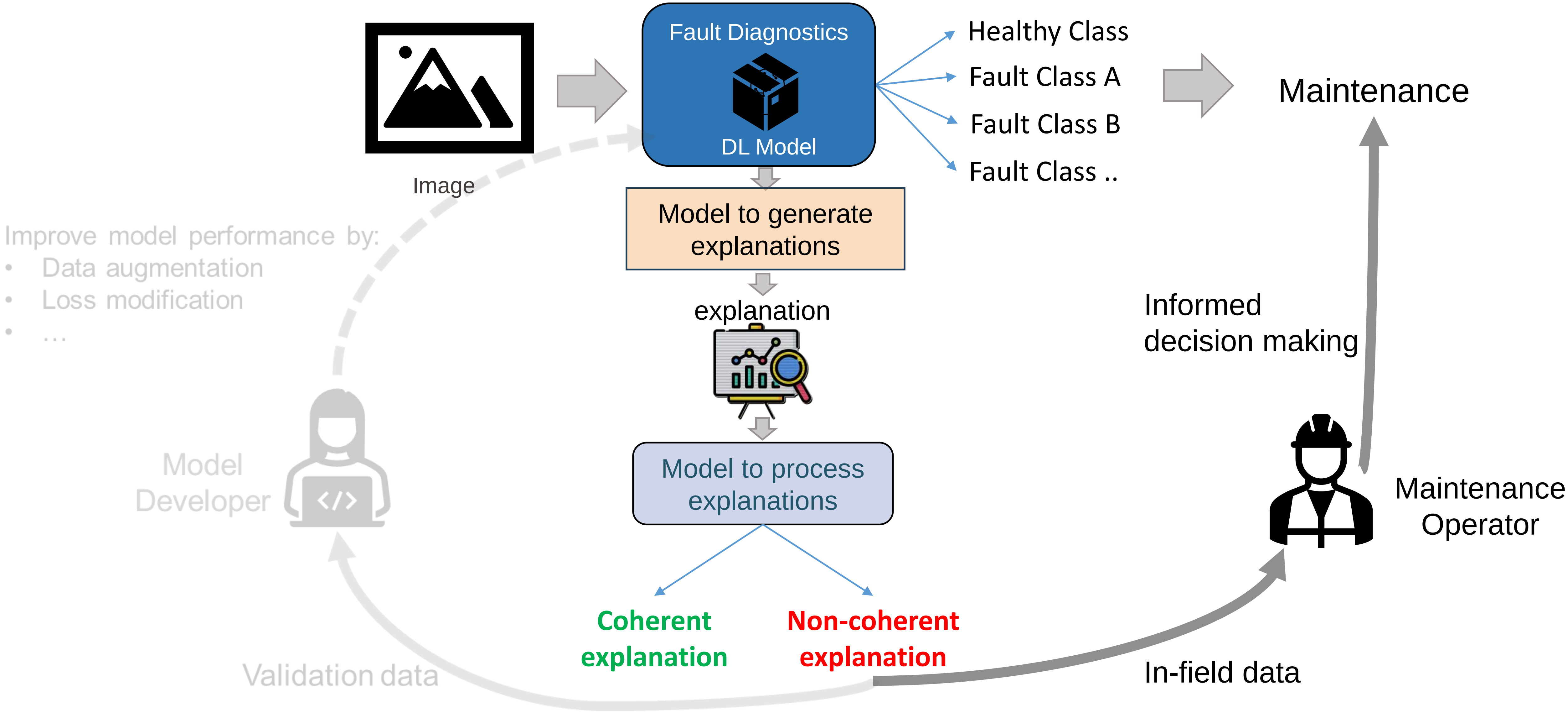
Relevance: Why is it Useful? (1)

Improve model performance by:

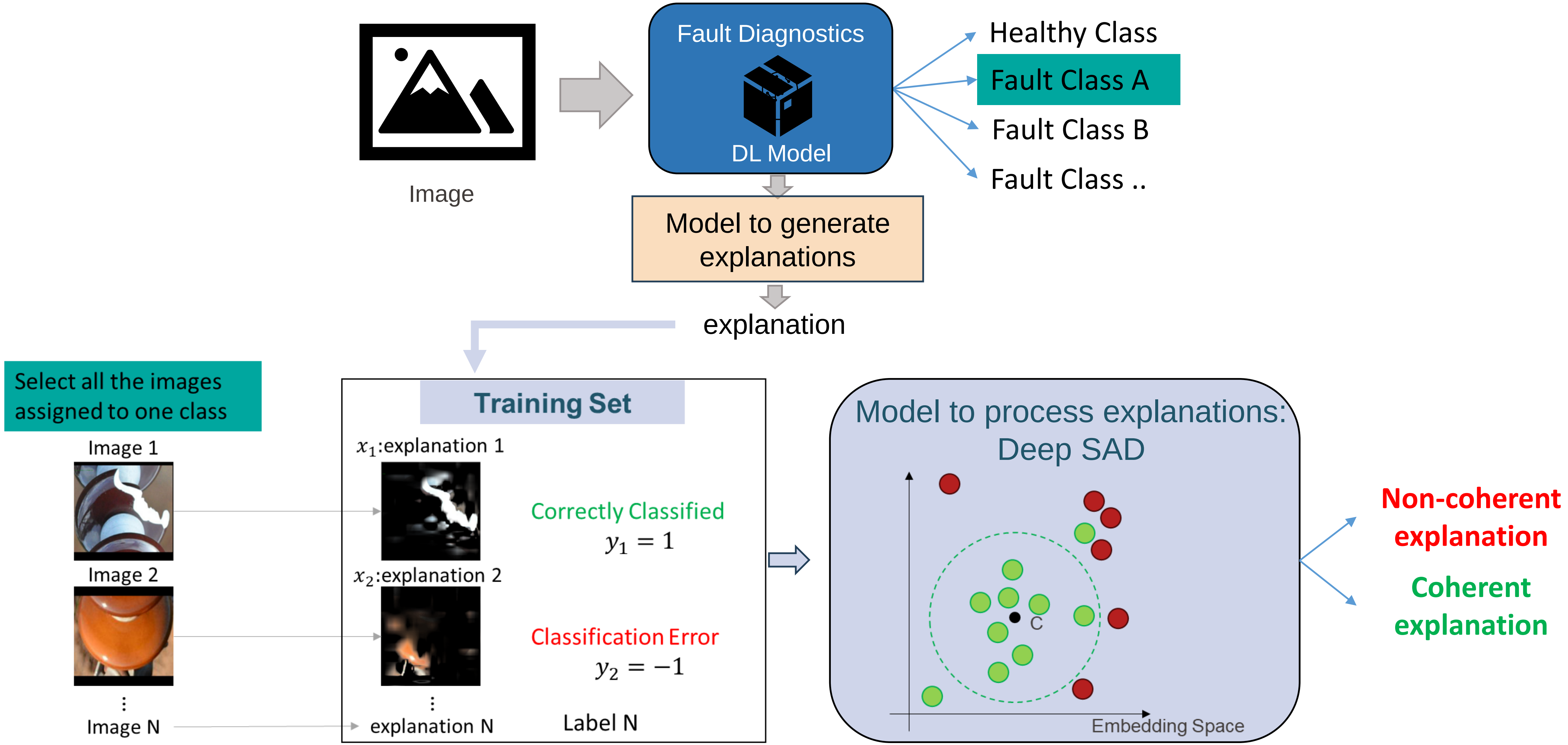
- Data augmentation
- Loss modification
- …



Relevance: Why is it Useful? (2)



Developed Method



Image

Fault Diagnostics

DL Model

Healthy Class

Fault Class A

Fault Class B

Fault Class ..

Model to generate explanations

explanation

Select all the images assigned to one class

Image 1

Image 2

⋮

Image N

Training Set

x_1:explanation 1

Correctly Classified

$y_1 = 1$

x_2:explanation 2

Classification Error

$y_2 = -1$

⋮

explanation N

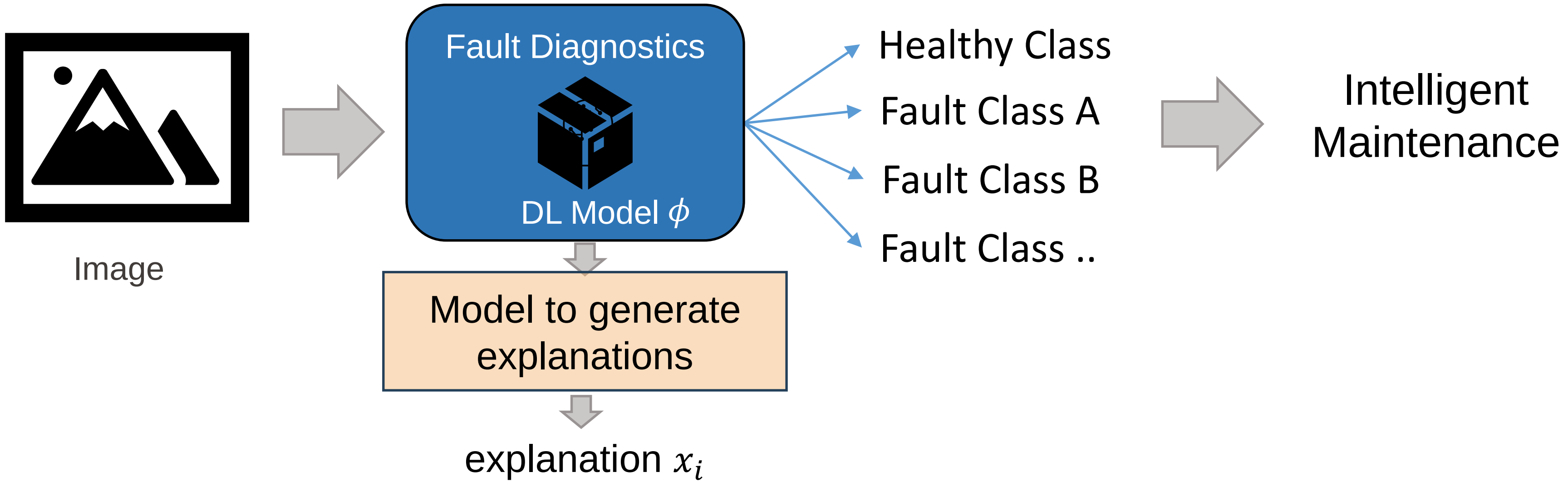
Label N

Model to process explanations: Deep SAD

Non-coherent explanation

Coherent explanation

Developed Method



Image

Fault Diagnostics

DL Model ϕ

Model to generate explanations

explanation x_i

Healthy Class

Fault Class A

Fault Class B

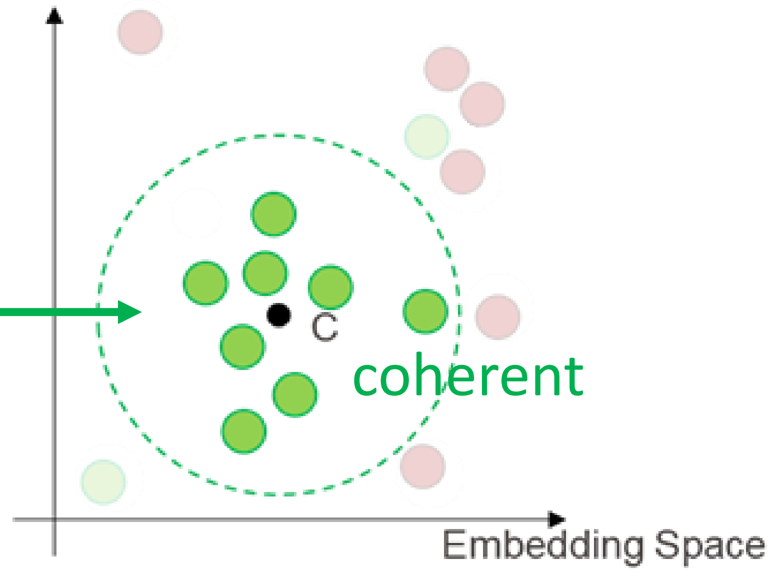
Fault Class ..

Intelligent Maintenance

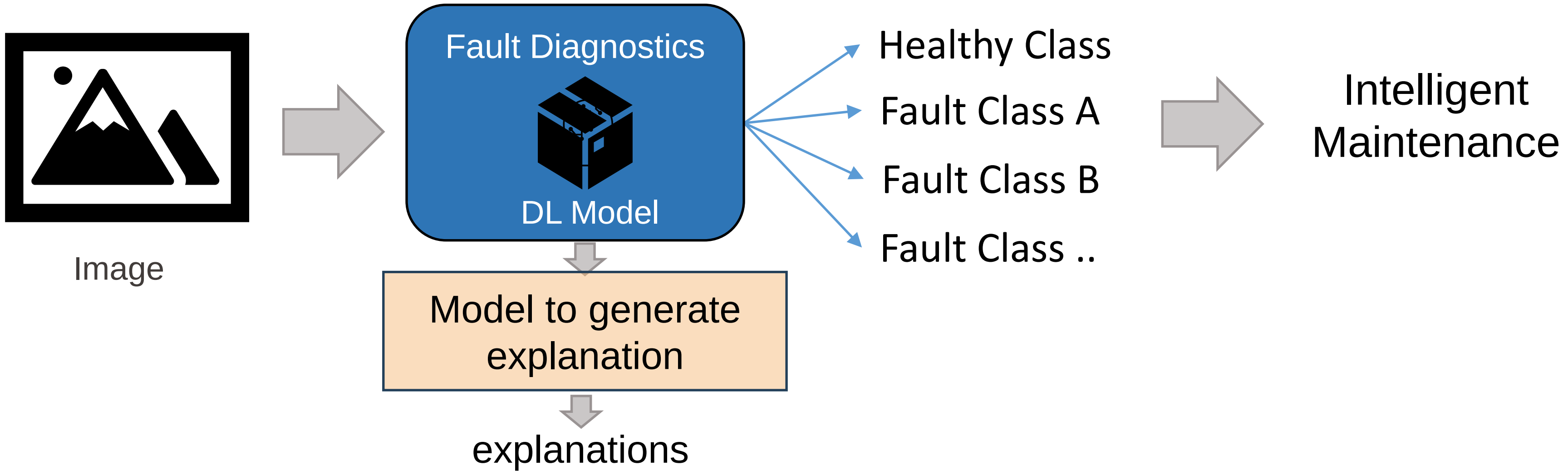
Embedding space definition: loss function

- Correct classifications: minimize the distance from the centre

$$\min_{\mathcal{W}} \quad \frac{1}{n+m}\sum_{i=1}^{n}\|\phi(x_i;\mathcal{W})-c\|^2$$



Developed Method



Image

Fault Diagnostics

DL Model

Healthy Class

Fault Class A

Fault Class B

Fault Class ..

Intelligent Maintenance

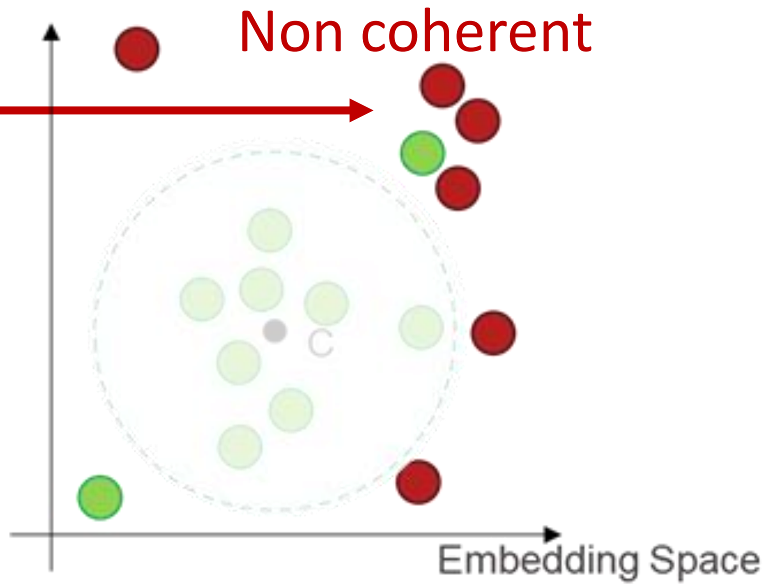
Model to generate explanation

explanations

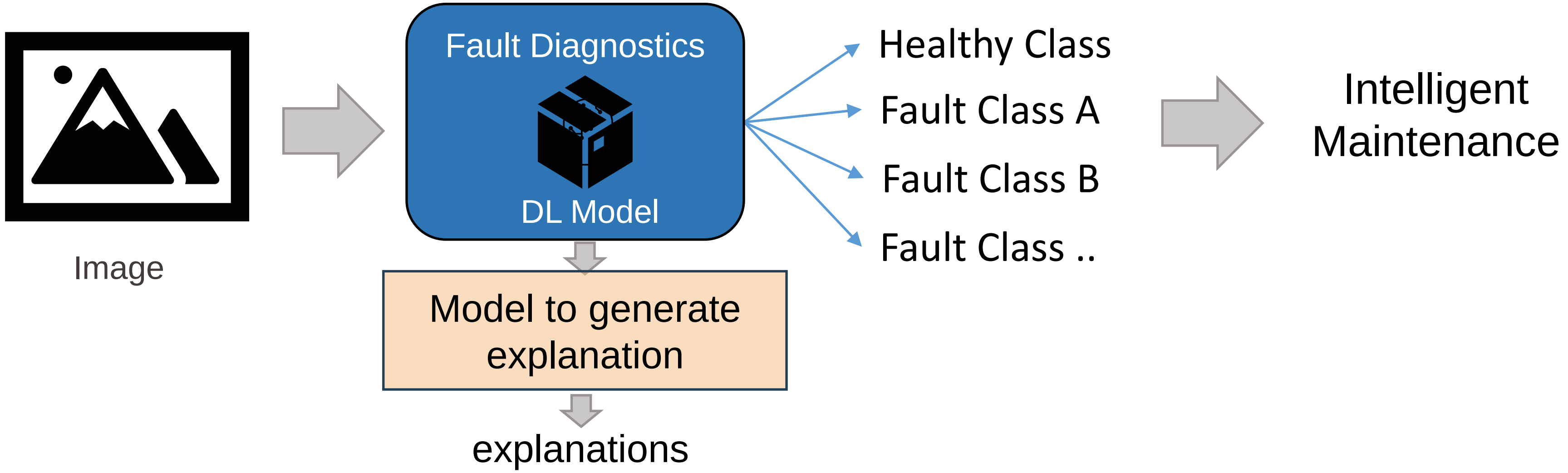
Embedding space definition: loss function

● Classifications errors:
maximize the distance from the centre
(hyperparameter η)

$$\min_{\mathcal{W}} \quad \frac{1}{n+m}\sum_{i=1}^{n}\|\phi(x_i;\mathcal{W})-c\|^2+\frac{\eta}{n+m}\sum_{j=1}^{m}\left(\|\phi(\tilde{x}_j;\mathcal{W})-c\|^2\right)^{\tilde{y}_j}.$$



Developed Method



Image

Fault Diagnostics

DL Model

Model to generate explanation

explanations

Healthy Class

Fault Class A

Fault Class B

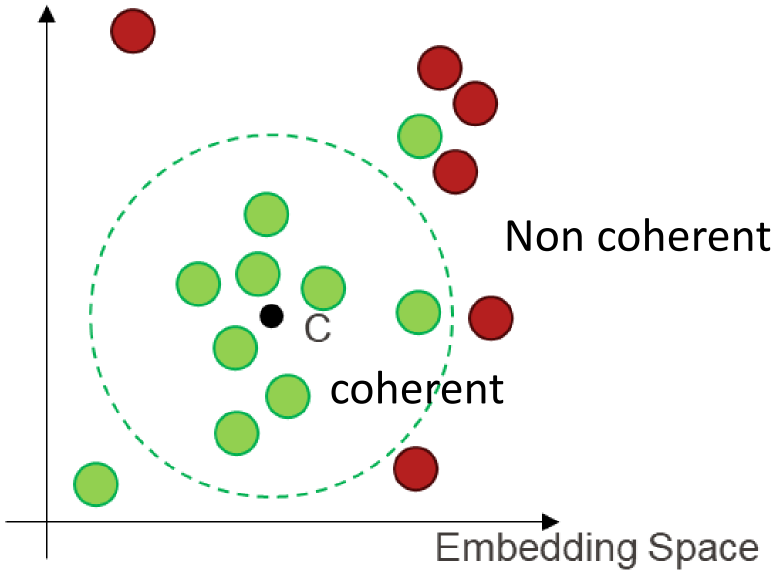
Fault Class ..

Intelligent Maintenance

Embedding space definition: loss function

Regularization term to avoid overfitting (hyperparameter λ)

$$\min_{\mathcal{W}} \quad \frac{1}{n+m}\sum_{i=1}^{n}\|\phi(x_i;\mathcal{W})-c\|^2+\frac{\eta}{n+m}\sum_{j=1}^{m}\left(\|\phi(\tilde{x}_j;\mathcal{W})-c\|^2\right)^{\tilde{y}_j}+\frac{\lambda}{2}\sum_{\ell=1}^{L}\|W^{\ell}\|_F^2.$$

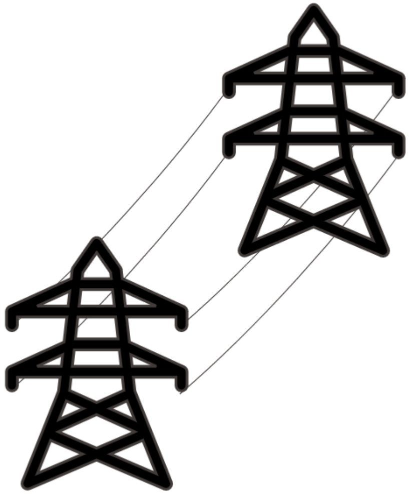


Case Study

Insulator Defect Image Dataset (IDID)

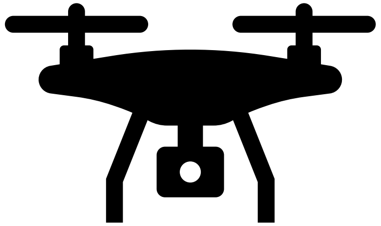
Dexter Lewis, Pratik Kulkarni, August 11, 2021, "Insulator Defect Detection", IEEE Dataport, doi: https://dx.doi.org/10.21227/vkdw-x769.s

Critical Components:

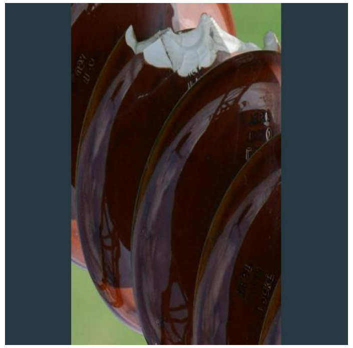


Power Grid

Insulators



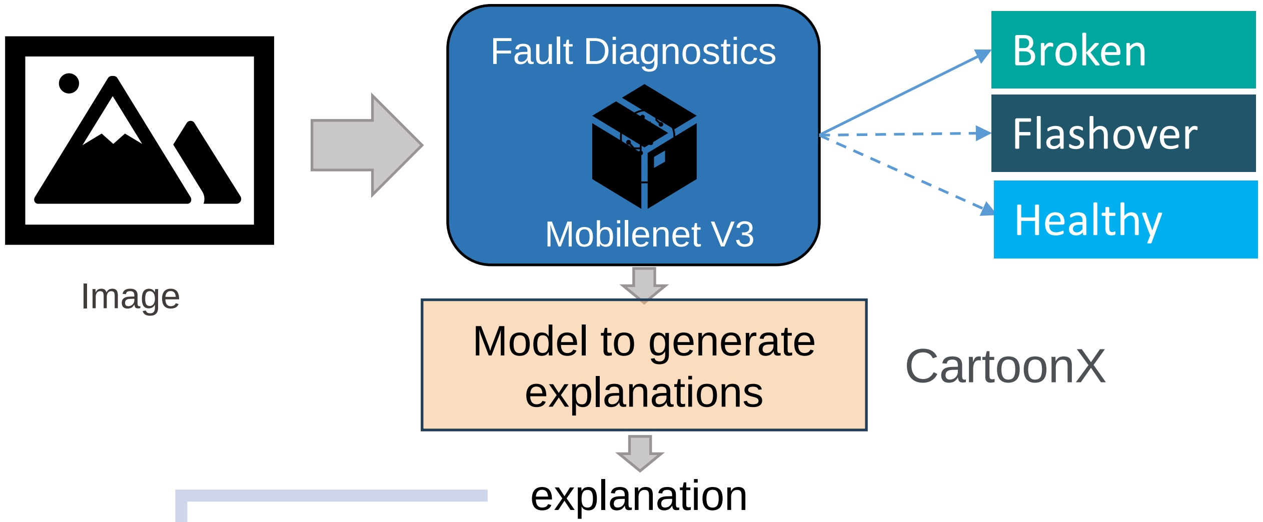
Drones



Shells’ images

e.g. swiss power grid:
6700 km long → 12 000 pylons

Case Study



Image

Fault Diagnostics

Mobilenet V3

Broken

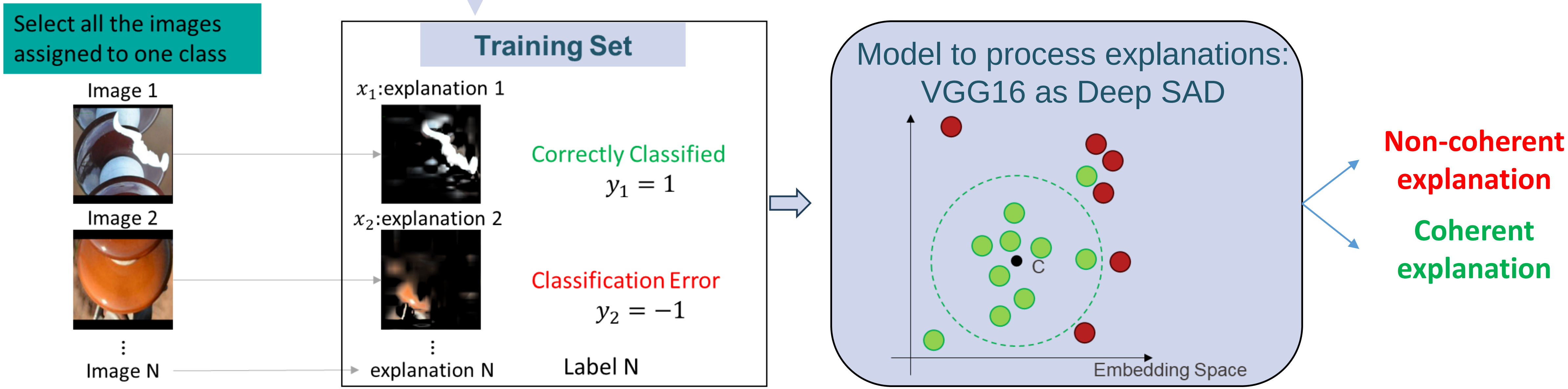
Flashover

Healthy

Model to generate explanations

CartoonX

explanation



Select all the images assigned to one class

Image 1

Image 2

⋮

Image N

Training Set

x_1:explanation 1

Correctly Classified

$y_1 = 1$

x_2:explanation 2

Classification Error

$y_2 = -1$

⋮

explanation N

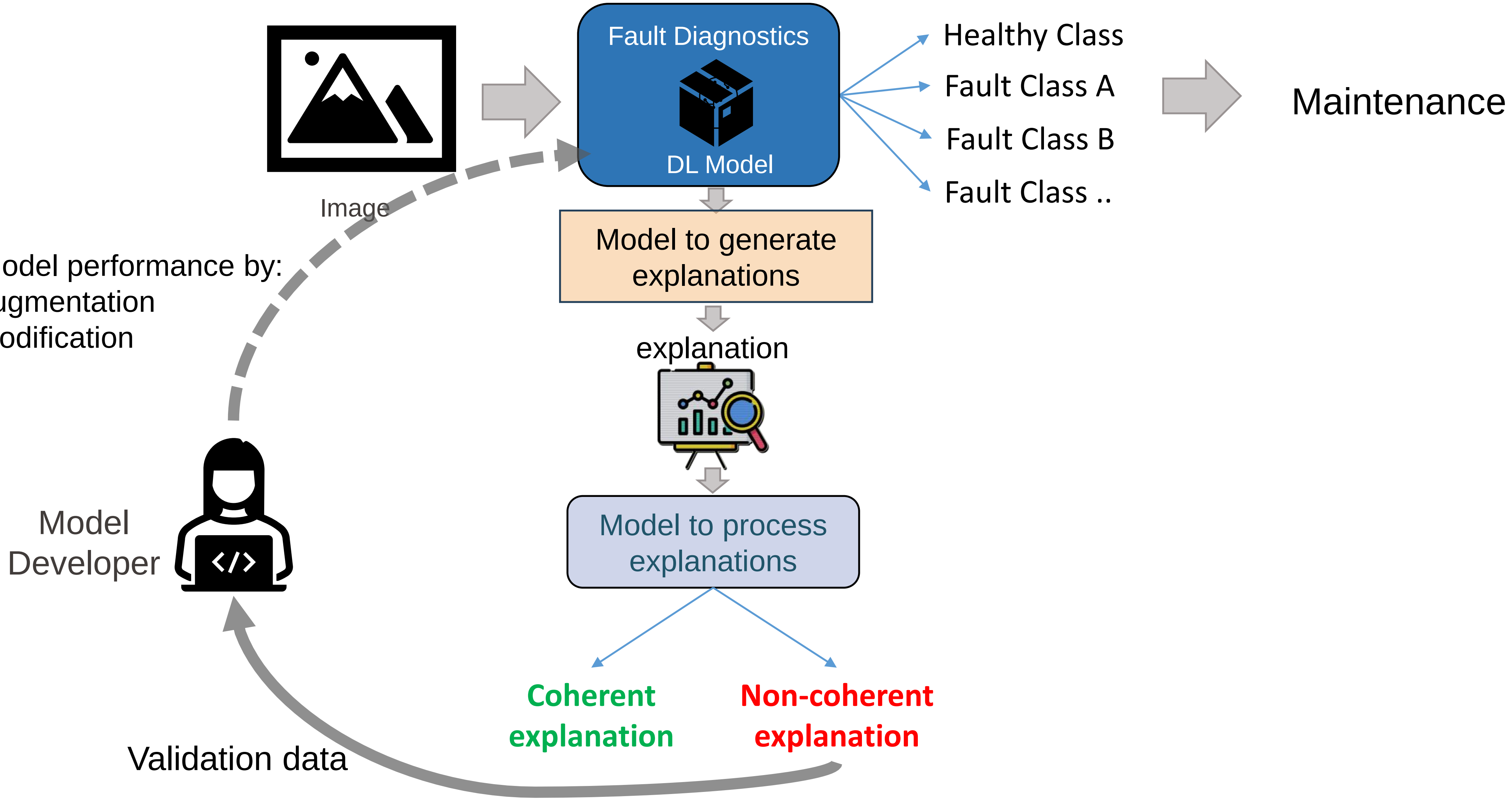
Label N

Model to process explanations: VGG16 as Deep SAD

Non-coherent explanation

Coherent explanation

Relevance: Why is it Useful? (1)



Image

Fault Diagnostics

DL Model

Healthy Class

Fault Class A

Fault Class B

Fault Class ..

Maintenance

Model to generate explanations

explanation

Model to process explanations

Coherent explanation

Non-coherent explanation

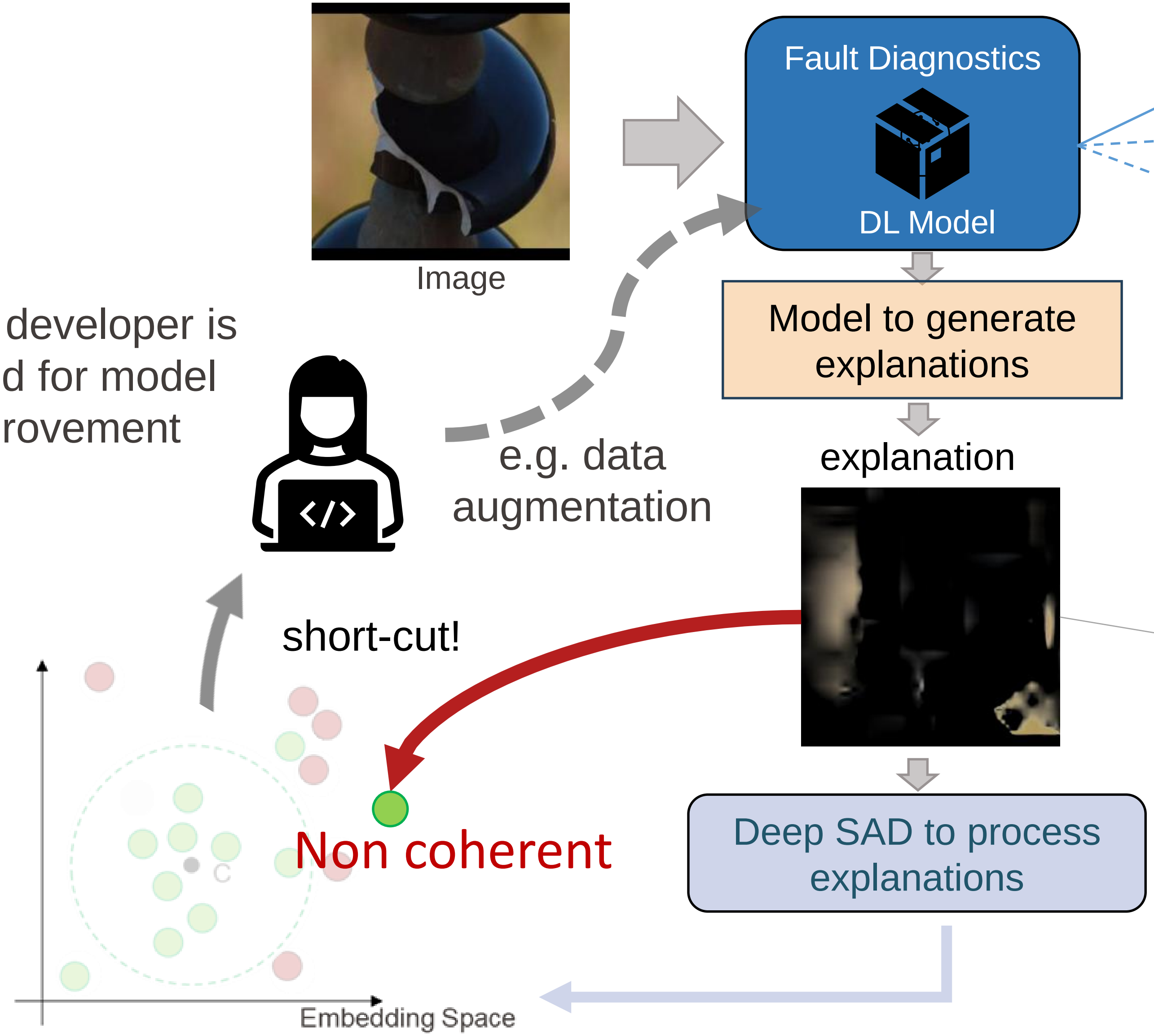
Validation data

Model Developer

Improve model performance by:

- Data augmentation
- Loss modification
- …

Results: Proposed Approach on Validation Set



Image

Fault Diagnostics

DL Model

Broken

Correct!

Flashover

Healthy

Model to generate explanations

explanation

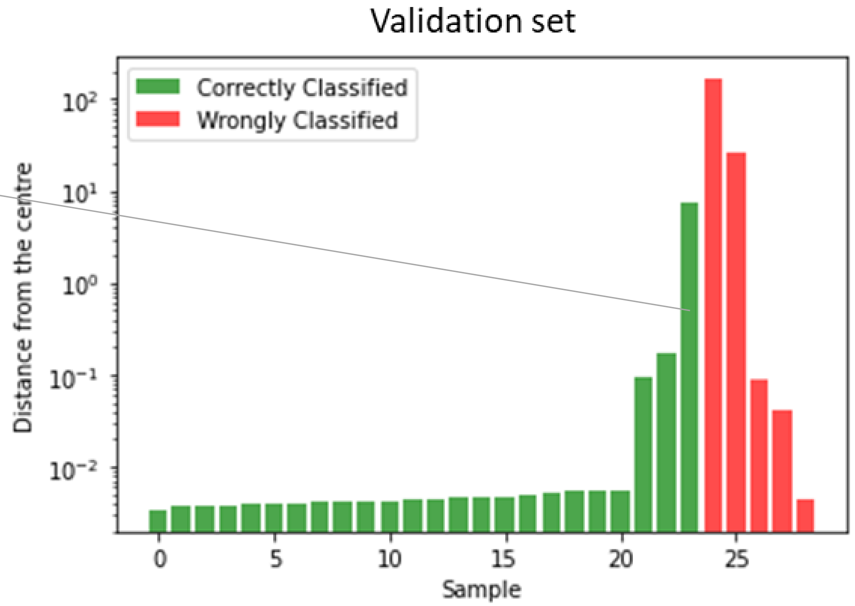
Deep SAD to process explanations

Model developer is guided for model improvement

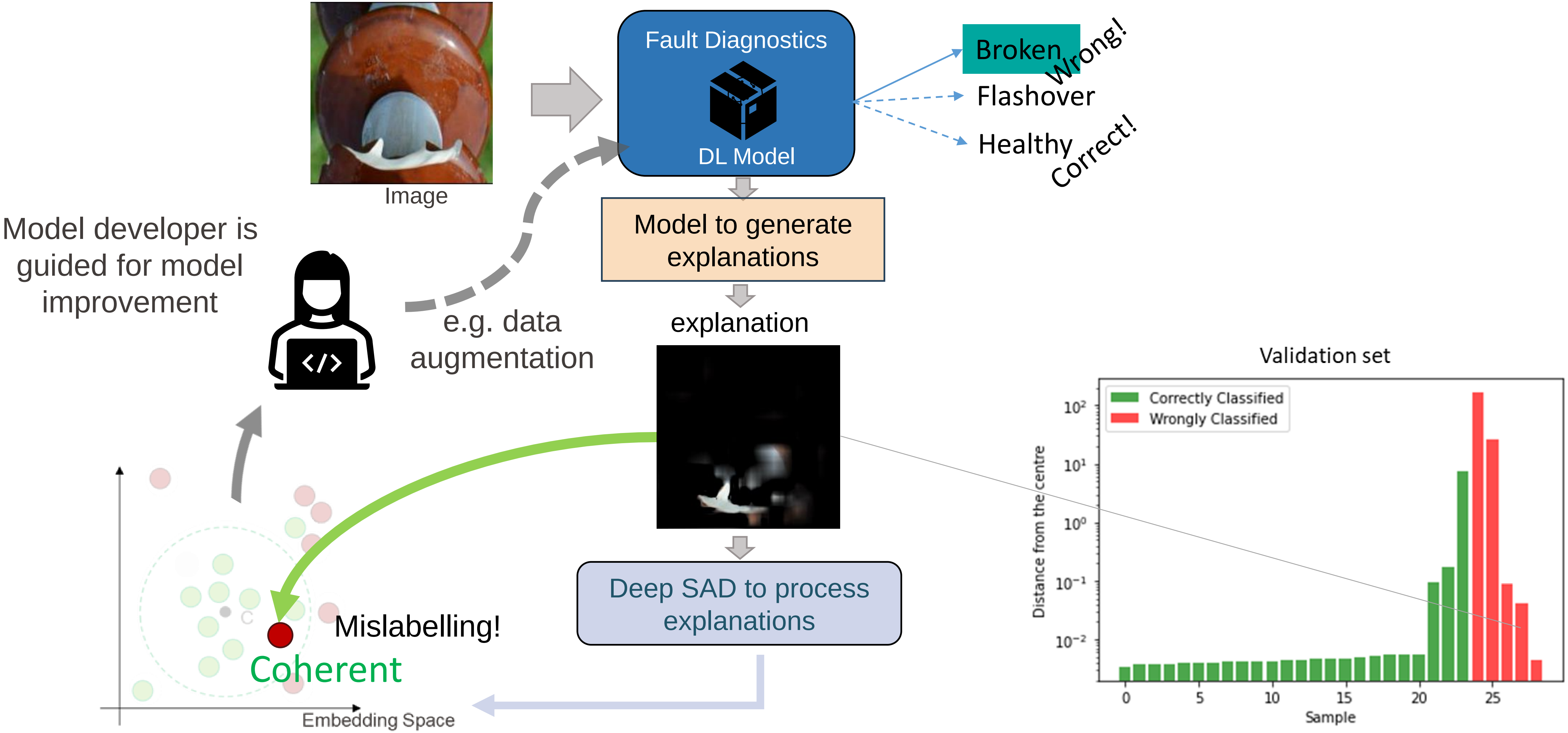
e.g. data augmentation

short-cut!

Non coherent



Results: Proposed Approach on Validation Set



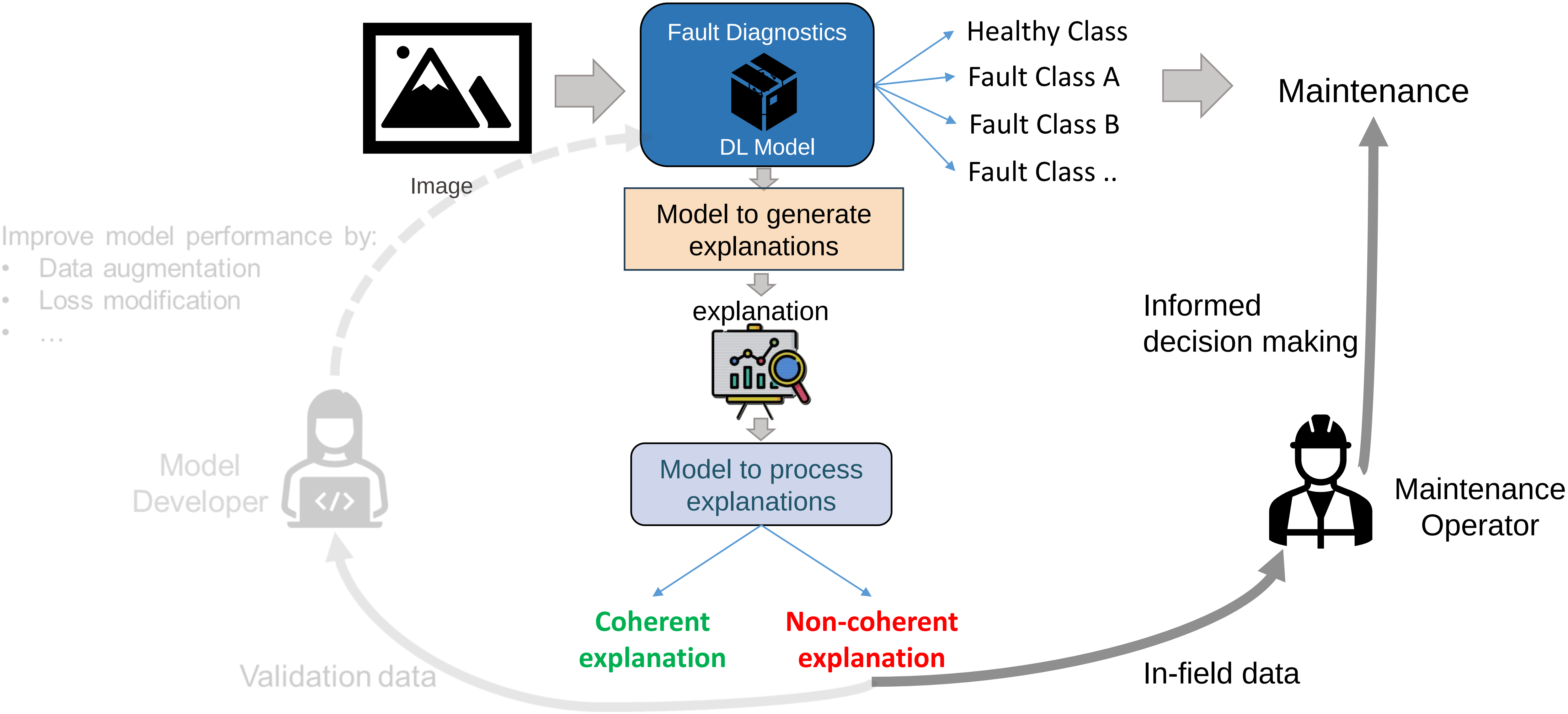
Model developer is guided for model improvement

e.g. data augmentation

Mislabelling!

Coherent

Relevance: Why is it Useful? (2)



Image

Fault Diagnostics

DL Model

Healthy Class

Fault Class A

Fault Class B

Fault Class ..

Maintenance

Model to generate explanations

explanation

Model to process explanations

Coherent explanation

Non-coherent explanation

Informed decision making

Maintenance Operator

In-field data

Improve model performance by:

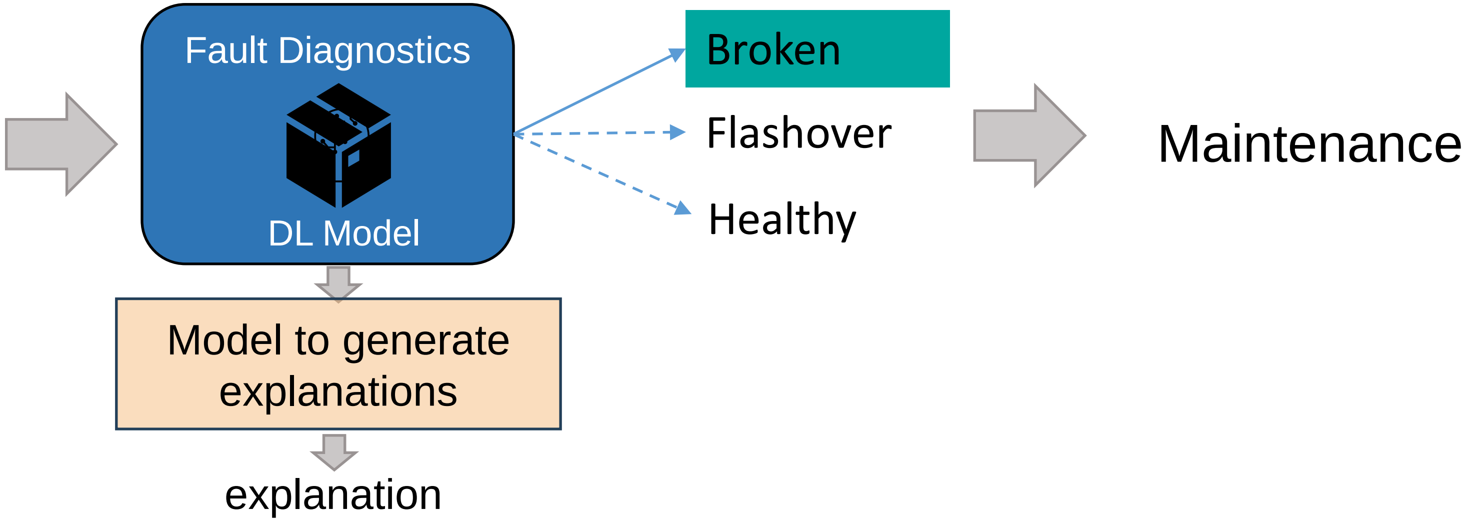
- Data augmentation
- Loss modification
- …

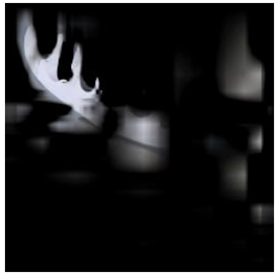
Model Developer

Validation data

Results: DL model

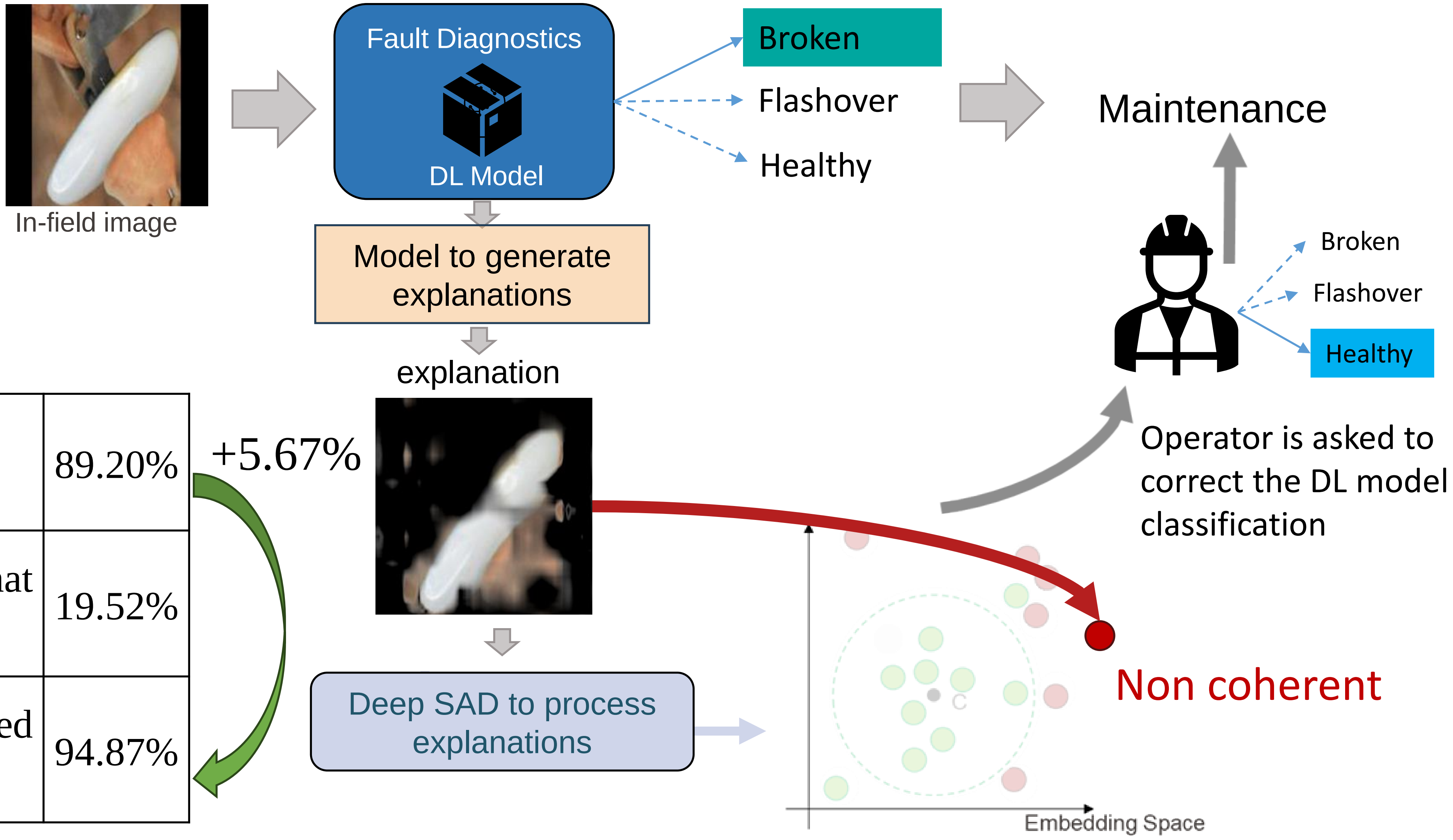
In-field image





Accuracy of DL model	89.20%

Results: Proposed Approach In-field Application



In-field image

Fault Diagnostics

DL Model

Broken

Flashover

Healthy

Maintenance

Model to generate explanations

explanation

Deep SAD to process explanations

Embedding Space

Non coherent

Broken

Flashover

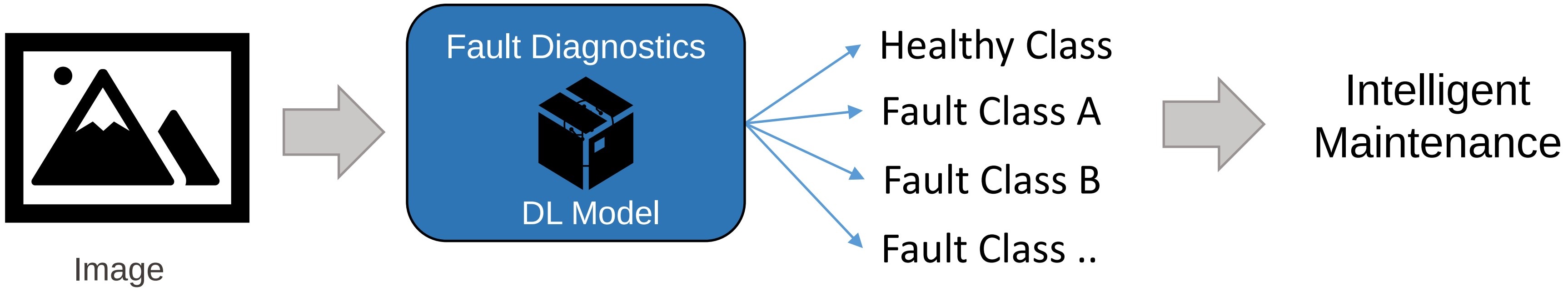
Healthy

Operator is asked to correct the DL model classification

+5.67%

Accuracy of DL model	89.20%
Fraction of explanations that is revised by experts	19.52%
Accuracy of the proposed approach	94.87%

Conclusion



Image

Fault Diagnostics

DL Model

Healthy Class

Fault Class A

Fault Class B

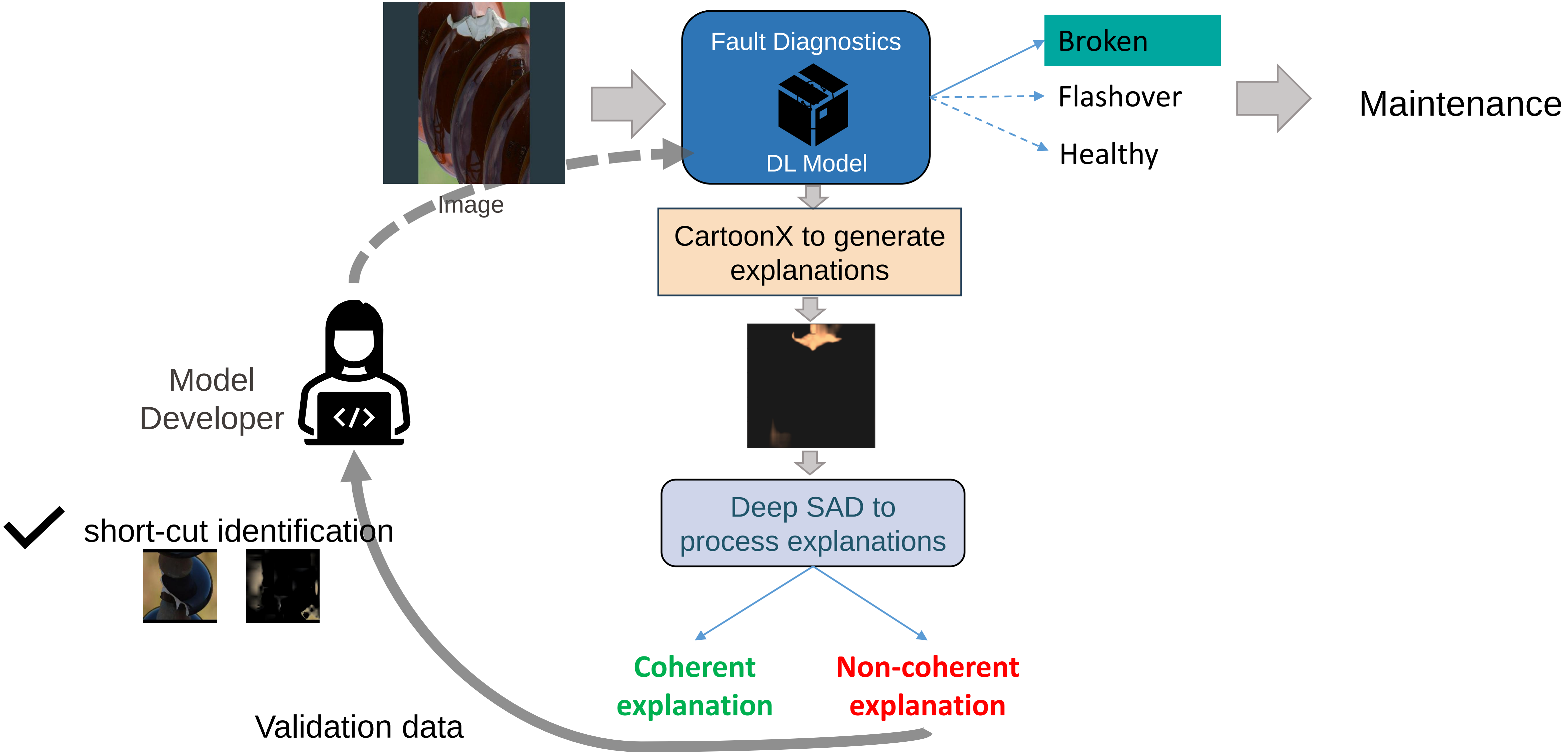
Fault Class ..

Intelligent Maintenance

Challenges:

- Black Box
- In-field performance drop

Conclusion



Image

Fault Diagnostics

DL Model

Broken

Flashover

Healthy

Maintenance

CartoonX to generate explanations

Deep SAD to process explanations

Coherent explanation

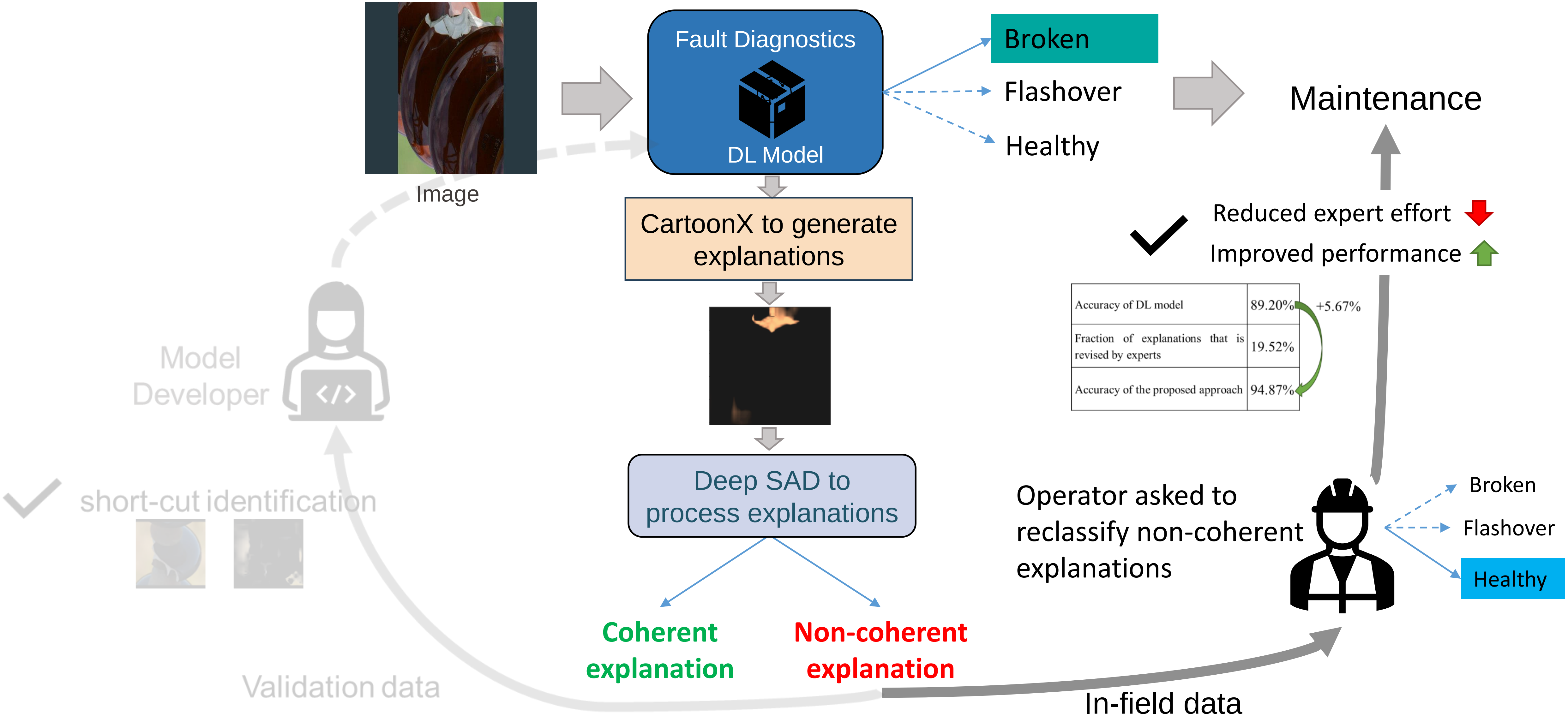
Non-coherent explanation

Validation data

Model Developer

✓ short-cut identification

Conclusion



Image

Fault Diagnostics

DL Model

Broken

Flashover

Healthy

Maintenance

CartoonX to generate explanations

Deep SAD to process explanations

Coherent explanation

Non-coherent explanation

Reduced expert effort

Improved performance

Accuracy of DL model	89.20%
Fraction of explanations that is revised by experts	19.52%
Accuracy of the proposed approach	94.87%

+5.67%

Operator asked to reclassify non-coherent explanations

Broken

Flashover

Healthy

In-field data

Model Developer

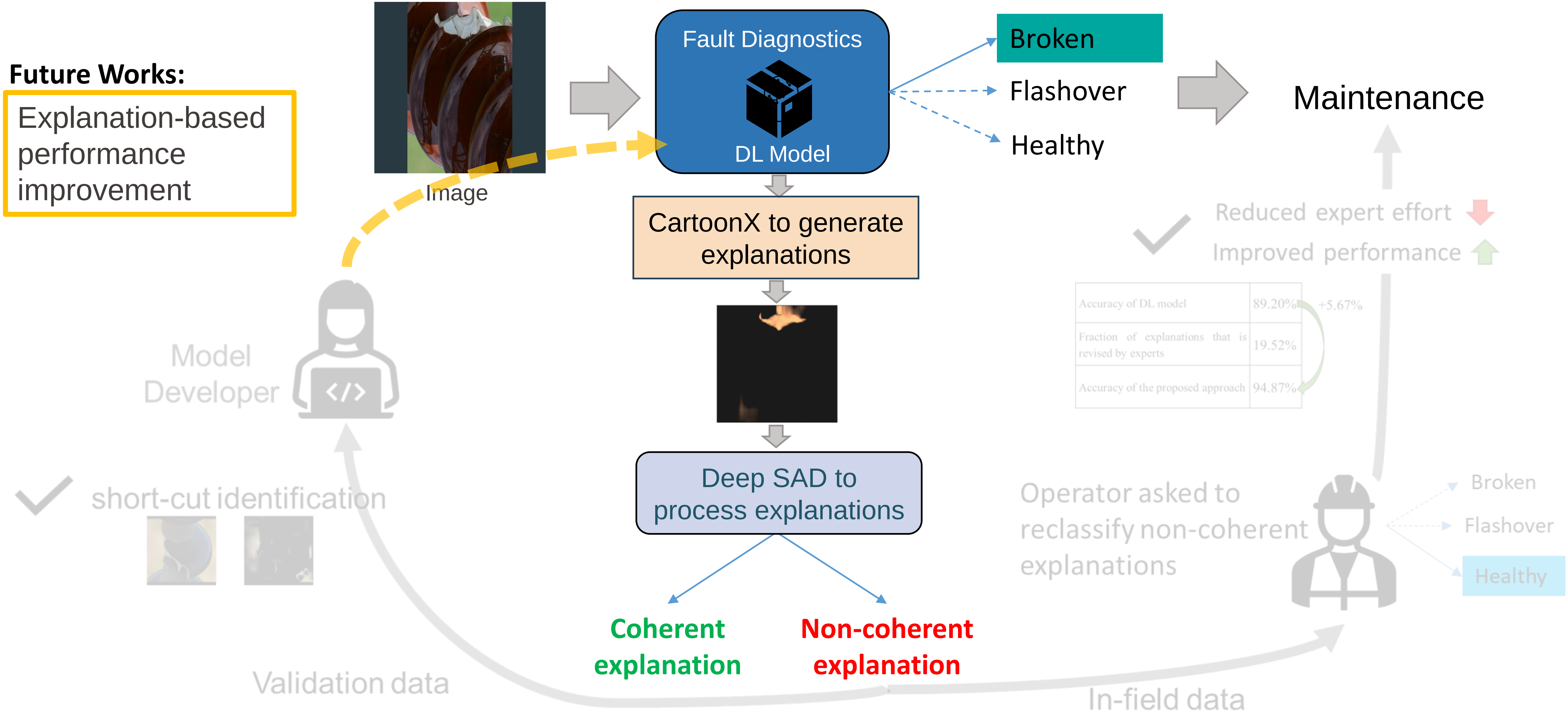
short-cut identification

Validation data

Conclusion

Future Works:

Explanation-based performance improvement



Image

Fault Diagnostics

DL Model

Broken

Flashover

Healthy

Maintenance

CartoonX to generate explanations

Deep SAD to process explanations

Coherent explanation

Non-coherent explanation

Model Developer

✓ short-cut identification

Validation data

In-field data

Operator asked to reclassify non-coherent explanations

✓ Reduced expert effort

Improved performance

Accuracy of DL model	89.20%	+5.67%
Fraction of explanations that is revised by experts	19.52%	
Accuracy of the proposed approach	94.87%	