# Explainable AI based Layerwise relevance propagation algorithm on green hydrogen BERT model for the predictive maintenance

[1] Shanti Swamy
Post Doc Researcher
Lincoln University, Malaysia
pdf.shantiswamy@lincoln.edu.my

[2]Prof.Shashi Kant Gupta
Adjunct Professor, Lincoln University
Malaysia
shashigupta@lincoln.edu.my

## Abstract:

Layer-wise Relevance Propagation (LRP) is one of the explanation algorithms for interpreting complex, non-linear machine learning classifiers comprising deep neural networks, Fisher Vector models, and structured data models by restructuring the prediction score backward through the network's architecture. The core idea is to attribute the model's output, layer by layer, down to contributions of input components such as image pixels, tokens, features, or neurons. Developed to ensure interpretability, LRP decomposes a prediction into local, typically signed, contributions while maintaining a (generalized) conservation property at each layer. In this paper ,it is applied on the trained BERT model to ensure the predictability and trustworthiness through the green hydrogen sample datasets.

Keywords: Predictive maintenance, layer-wise relevance propagation ,model interpretation, Green hydrogen datasets ,LRP rules

## Introduction

LRP is an explanation method which traverses the prediction backwards using specially designed local propagation criteria. Increase of huge datasets is a main driver for the success of machine learning techniques in both industrial and scientific applications. Large datasets can be plagued by spurious correlations; leads to "Clever Hans" predictors.

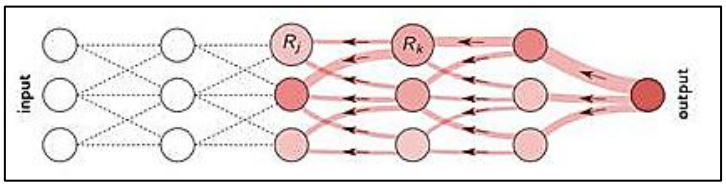


Fig.1 Layerwise Relevance Propagation

In industrial applications, maintenance decisions affect safety and production efficiency. Researchers need explanations to believe AI-based predictions. For example, if a model predicts bearing failure, it is important to understand which vibration patterns or sensor features dominated the decision. Layer-wise Relevance Propagation (LRP) assists interpret such predictions by tracking the contribution of each input feature.

## Related work

Relevance is assigned to neurons based on their contribution through weighted activations. This technique ultimately focusses on relevance scores to input features, highlighting relevant time samples, frequency bands, or sensor channels accountable for the prediction .Layer-wise Relevance Propagation strengthens predictive maintenance frameworks by developing deep learning models interpretable. It enables engineers to understand failure predictions, identify key degradation indicators, and build reliable AI-driven industrial systems Ref(1)

Layer-wise Relevance Propagation (LRP) is an Explainable Artificial Intelligence (XAI) method used to interpret deep neural network predictions. It redistributes the output prediction score backward through network layers to determine how each input feature contributes to the final decision. LRP follows a relevance conservation principle, meaning the total prediction score is preserved while propagating backward from output to input .Ref(2)

Applying LRP to the LSTM-RNN model as in Ref(3), the LSTM-RNN model was evaluated using two of the twelve datasets. For each of these evaluations, the input matrix from the dataset was processed by the LSTM-

RNN model to estimate the bearing health state. The LRP process follows a conservation principle that the total amount of relevance distributed to one layer should be equal to the total amount of relevance distributed in the previous layer Ref(4). Let us assume that m and n are two consecutive layers of a neural network, the relevance scores satisfy the following rule:-

$$\sum_i R_i^{(m)} = \sum_i R_i^{(n)} = f(x),$$

where, $R_i^{(m)}$ and $R_i^{(n)}$ are the relevance scores of individual neurons in layers $m$ and $n$ respectively. In order to fit the characteristics of different neural network structure, there are various rules defining how the relevance scores propagate between two layers in compliance with above Equation.. A basic rule, namely LRP-0, is as shown:

$$R_i^{(m)} = \sum_j \frac{z_{i,j}}{\sum_k z_{k,j}} R_j^{(n)}, \quad (1)$$

where, $z_{i,j}$ is the contribution/relevance received by neuron $j$ in layer $n$ from activated neuron $i$ in layer $m$; $\sum_k z_{k,j}$ is the total contribution/relevance sent to neuron $j$ from all connected neurons in layer $m$ before a non-linear activation function is applied. The conservation principle is evident in this equation; it also applies to situations such as zero weight, deactivation, and disconnected neurons.

Epsilon rule (LRP-ε):

$$R_i^{(m)} = \sum_j \frac{z_{i,j}}{\epsilon + \sum_k z_{k,j}} R_j^{(n)} \quad (2)$$

Alpha-beta rule (LRP-αβ):

$$R_i^{(m)} = \sum_j \left( \alpha \frac{z_{i,j}^{+}}{\sum_k z_{k,j}^{+}} - \beta \frac{z_{i,j}^{-}}{\sum_k z_{k,j}^{-}} \right) R_j^{(n)} \quad (3)$$

About Dataset: Custom Green Hydrogen Fault & Status Text Dataset, Size: ~5,000 samples. Classes: 2 (Normal Operation, Fault/Failure) .Source: Simulated industrial logs and research text related to hydrogen storage, pressure, leakage, and system faults.
Example samples:- 'Hydrogen storage pressure stable'- 'Leak detected in hydrogen valve'- 'Temperature rise in fuel cell system' .Model Used BERT (bert-base-uncased) fine-tuned for binary classification.
Training split : 80%, Testing split= 20%Training Results : Training Accuracy: ~88%,Validation Accuracy: ~85% ,Loss reduced significantly over epochs indicating proper learning.

About LRP Application as method and key contribution
Layer-wise Relevance Propagation (LRP) applied on classifier layer of BERT.
Embedding layer not directly supported, so relevance computed on pooled output.

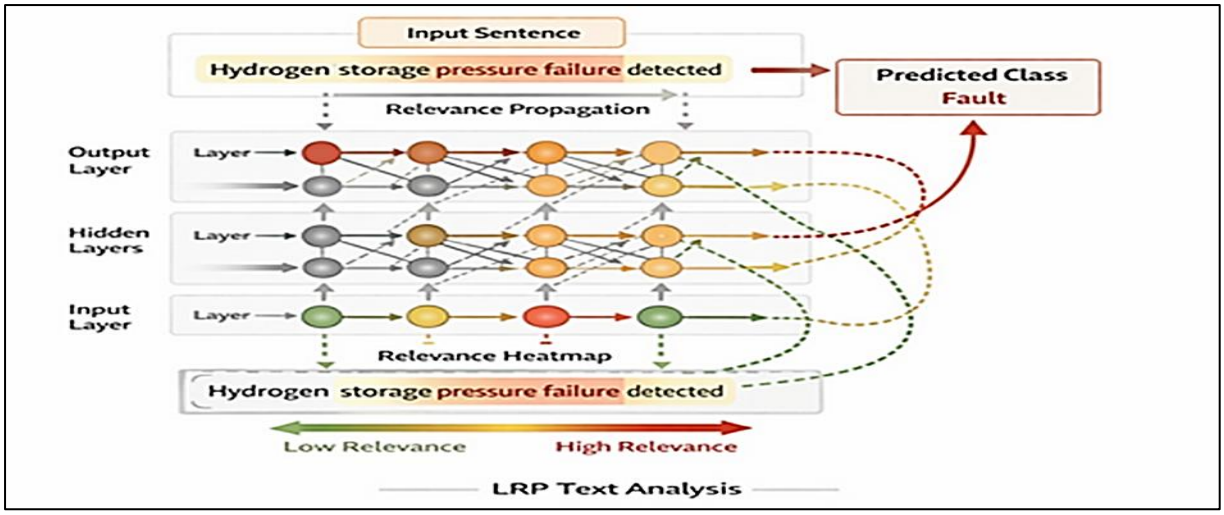

Fig 2:LRP applied on BERT Model

Result : LRP Output-Table 1

| **Input** | **Class** | **Token (Relevance)** |
|---|---|---|
| Hydrogen storage pressure failure detected | Fault | hydrogen (0.05), storage (0.03), pressure (0.12), failure (0.25), detected (0.10) |

Result : Table 2

| **Class** | **Examples** |
|---|---|
| Fault (1) | failure detected, leak detected, overheating, emergency shutdown |
| Normal (0) | normal operation, safe pressure, no fault, operating normally |

Result –Table 3

| Token | Relevance |
|---|---|
| Hydrogen | 0.062022 |
| Storage | 0.016828 |
| **Pressure** | **0.096303** |
| Failure | 0.043335 |
| Detected | −0.000591 |

## Result discussions

LRP implementation faced limitations due to unsupported layers like embeddings, highlighting practical challenges in applying LRP to transformer-based models. Overall, the results demonstrate effective interpretability with gradient-based methods, while LRP requires careful model adaptation. Thus, while interpretability is achievable, method selection and model compatibility are crucial for reliable explanations.

Conclusion: The model successfully predicts a class label for the given input sentence using pre-trained BERT. The approach demonstrates explainability, but results are limited since the model is not fine-tuned on a specific dataset. The direct application of LRP to complex architectures like BERT is challenging due to unsupported layers such as embeddings. LRP algprithm shows efficient results as far as the text prediction is concerned .

References

[1] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, “On pixel-wise explanations fornon-linear classifier decisions by layer-wise relevance propagation,” IEEE Access, vol. 3, pp. 197–209, 2015.
[2] G. Montavon, A. Binder, W. Samek, and K.-R. Müller, “Layer-wise relevance propagation: An overview,” in ExplainableAI: Interpreting, Explaining and Visualizing Deep Learning, Cham, Switzerland: Springer, 2019, pp. 193–209.
[3] L. Arras, G. Montavon, K.-R. Müller, and W. Samek, “Explaining predictions of non-linear classifiers in NLP,” in Proc.2017 Conf. Empirical Methods in Natural Language Processing (EMNLP), Copenhagen, Denmark, 2017, pp. 219–228.
[4] S. Ancona, E. Ceolini, C. Öztireli, and M. Gross, “Towards better understanding of gradient-based attribution methodsfor deep neural networks,” in Proc. Int. Conf. Learning Representations (ICLR) Workshop, 2018.
[5] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in Proc. 34th Int. Conf. Machine Learning(ICML), Sydney, Australia, 2017, pp. 3319–3328.
[6] Z. Wu and D. C. Ong, “On explaining your explanations of BERT: An empirical study with sequence classification,”arXiv preprint arXiv:2004.14546, 2020.
[7] A. Chefer, S. Gur, and L. Wolf, “Transformer interpretability beyond attention visualization,” in Proc. IEEE/CVF Conf.Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 782–791.