Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p27773-cd

TCAV-Based Operator Support System for Human Reliability Enhancement in Safety-Critical Environments

Young Ho Chae

Advanced Instrumentation and Control Division, Korea Atomic Energy Research Institute, Republic of Korea. E-mail: yhchae@kaeri.re.kr

Seo Ryong Koo

Advanced Instrumentation and Control Division, Korea Atomic Energy Research Institute, Republic of Korea. E-mail: srkoo@kaeri.re.kr

This study introduces a concept-level explanation method for AI-based nuclear power plant (NPP) diagnostic systems using Testing with Concept Activation Vectors (TCAV). Current explainable AI (XAI) methods, such as Layer-wise Relevance Propagation (LRP), provide sensor-level attributions that identify which input variables are important for the model's prediction. However, these attributions lack behavioral context—they indicate the importance of a sensor without describing how the sensor behaves. This limitation makes it difficult for operators to interpret and verify the AI's reasoning during complex diagnostic situations.
To address this limitation, we propose a TCAV-based operator support system that provides concept-level explanations. The proposed method defines behavioral concepts (e.g., "very high," "low," "oscillating") that correspond to recognizable sensor signal patterns in NPP operations. These concepts are combined with PageRank-weighted centrality scoring to rank the most influential and structurally important diagnostic patterns. We applied the proposed method to the IAEA iPWR simulator dataset, focusing on two representative scenarios: feedwater pump failure (AB_01) and loss of coolant accident (EM_01).
The experimental results demonstrate that LRP and TCAV identify fundamentally different aspects of model behavior, with only 5–10% overlap between the two methods. While LRP tends to highlight directly affected sensors (e.g., pump speed during a pump failure), TCAV identifies system-level behavioral patterns (e.g., steam generator inlet temperature being very high) that provide more informative reasoning for operators. The proposed approach offers a practical framework for generating AI explanations that operators can directly verify against their domain knowledge and current plant observations.

Keywords: Explainable AI, Test with Concept Activation Vector, Nuclear Power Plant, Operator Suppoprt, Concept-level Explanation.

1. Introduction

The application of artificial neural networks (ANNs) for nuclear power plant (NPP) diagnostics has been extensively investigated (Bae, Park, and Lee 2022, Chae et. al. 2025) in recent years. Deep learning models, particularly LSTM-based architectures, have demonstrated high classification accuracy for detecting and identifying abnormal and emergency scenarios from multivariate time-series data. However, a significant challenge remains in making these models' predictions interpretable to plant operators.

Explainable AI (XAI) methods have been widely applied to address this challenge. In our previous work (Chae et. al. 2025), we compared various gradient-based attribution methods—including LRP, DeepSHAP, Integrated Gradients, LIME, and saliency maps—for feature selection in NPP diagnostics. These methods successfully identified the most relevant sensors for distinguishing between different plant conditions. However, the outputs of these methods are inherently at the sensor level: they produce ranked lists of sensors with

numerical importance scores (e.g., "Pump speed 2: relevance 1.00"). While useful for feature selection and dimensionality reduction, this type of output has a limitation when used for operator decision support—it does not describe how the identified sensors behave.

For instance, when a feedwater pump fails, LRP identifies pump speed as the most important sensor. This is statistically valid, but operators already know that pump speed is directly affected during a pump failure. Additionally, operators need for complex diagnostic reasoning is information about downstream effects: whether steam generator temperatures are rising abnormally, whether coolant flow rates are declining, or whether thermal instabilities are developing in the fuel cladding. In other words, providing proper behavioral information about system-level changes is more useful for supporting operator reasoning than simply identifying which sensors have the largest attribution values.

Testing with Concept Activation Vectors (TCAV) (Kim, et. al. 2018) offers a different approach. Instead of attributing predictions to individual input features, TCAV tests whether high-level concepts—defined as directions in a neural network's activation space—influence the model's predictions. In the NPP context, these concepts can be defined as behavioral patterns of sensor signals: a temperature being "very high," a flow rate being "low," or a pressure signal "oscillating." These descriptors correspond to how operators are trained to describe and monitor system states.

This study presents a TCAV-based operator support system for NPP abnormal and emergency diagnosis. The key contributions of this study are as follows:

(i) We define a behavioral concept library consisting of seven temporal patterns (high, low, very high, very low, oscillating, step change, spike) that correspond to recognizable sensor signal behaviors in NPP operations.

(ii) We propose a PageRank-weighted concept ranking method that combines TCAV importance with structural centrality from sensor interaction graphs.

(iii) We provide a systematic comparison between LRP attribution and TCAV-based explanation on two distinct NPP fault scenarios, demonstrating that these methods capture fundamentally different aspects of model behavior.

(iv) We demonstrate that TCAV explanations provide behavioral context that operators can directly verify against their observations, whereas LRP provides only numerical attribution values.

2. Background

2.1. *Explainable AI for Time-Series Classification*

Gradient-based attribution methods have been widely applied to time-series classification models for identifying important input features. Given a trained model f and input $\mathrm{x} \in \mathbb{R}^{T \times D}$ (where T is the number of timesteps and D is the number of sensors), LRP computes relevance scores for each input element. In this work, we employ the Gradient × Input variant, which computes relevance as:

$$\boldsymbol{R_{t,d}} = \boldsymbol{x_{t,d}} \cdot \frac{\boldsymbol{\partial f}(\mathrm{x})}{\boldsymbol{\partial x_{t,d}}} \tag{1}$$

This produces a relevance map $\mathbf{R} \in \mathbb{R}^{T \times D}$ that can be aggregated across time to produce sensor-level importance rankings. While computationally efficient and applicable to any differentiable model, the resulting explanations are inherently feature-level: they identify important sensors without characterizing the nature of their importance.

Alternative approaches including SHAP, Integrated Gradients, and attention-based

explanations share this limitation. They identify which features are important but do not describe what behavioral patterns are important—a distinction that is relevant for operator interpretation.

2.2.Testing with Concept Activation Vectors

TCAV (Kim, et. al. 2018) provides concept-level explanations by learning directions in a neural network's internal activation space that correspond to human-defined concepts. Given a concept C (e.g., "temperature is very high"), a linear classifier—the Concept Activation Vector (CAV)—is trained to distinguish between activations produced by concept-positive and concept-negative examples. The TCAV score quantifies the sensitivity of the model's predictions to the learned concept direction:

$$\mathbf{TCAV}_C = \frac{|\{\boldsymbol{x} \in \boldsymbol{X_k}: \boldsymbol{S_C}(\boldsymbol{f_l}(\boldsymbol{x})) > \mathbf{0}\}|}{|\boldsymbol{X_k}|} \quad (2)$$

where S_C is the directional derivative of the model output along the CAV direction in the activation space of layer l, and X_k represents inputs from class k. A TCAV score of 1.0 indicates that the concept consistently drives predictions toward the target class; 0.5 indicates no influence.

The advantage of TCAV for operator support is that concepts can be defined to match how operators describe system states. Instead of "sensor 44 has relevance 1.0," TCAV can report "the concept of steam generator inlet temperature being very high strongly drives the fault classification."

3. Methodology

3.1.*System Architecture*

Fig. 1 illustrates the architecture of the proposed TCAV-based operator support system. The system consists of three components: (1) a diagnosis model that classifies plant operational scenarios from multivariate time-series data, (2) an LRP attribution module that serves as the baseline, and (3) a TCAV-based concept explanation module.

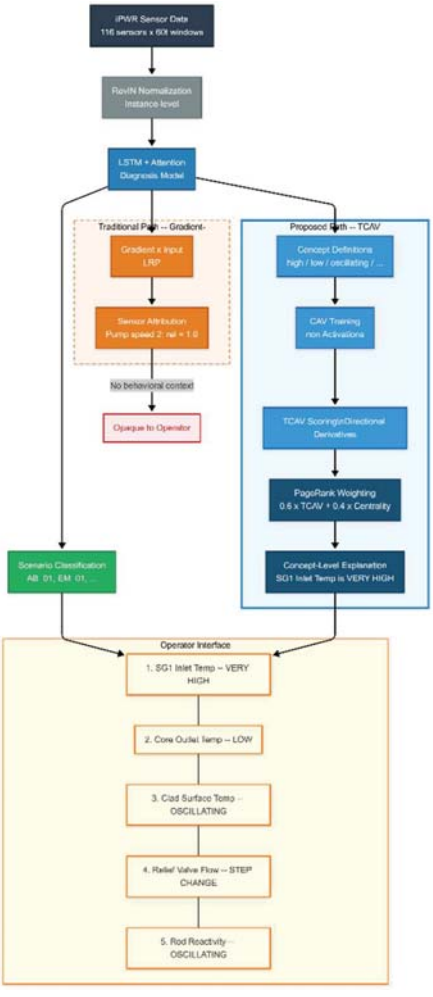

Fig. 1 System Architectures

3.2.*Diagnosis Model*

The diagnosis model processes multivariate time-series windows using a sequential architecture. Input windows $\mathbf{x} \in \mathbb{R}^{60\times116}$ (60 timesteps, 116 sensors) are first normalized using Reversible Instance Normalization (RevIN) (Kim et. al. 2022), which applies per-instance, per-sensor normalization:

$$\hat{\boldsymbol{x}}_{\boldsymbol{t,d}} = \frac{\boldsymbol{x_{t,d}} - \boldsymbol{\mu_d}}{\boldsymbol{\sigma_d} + \boldsymbol{\epsilon}} \quad (3)$$

where μ_d and σ_d are the mean and standard deviation of sensor d within the window. This

normalization handles distribution shifts across different operating conditions.

The normalized input is processed by a two-layer LSTM network with an attention mechanism. The attention layer weights timesteps within each window, enabling the model to focus on diagnostically informative time periods. The final classification is produced by a fully connected layer mapping the attended representation to one of 35 scenario classes (26 accident benchmarks, 9 emergency management scenarios).

The model specifications are as follows:

- Structure: LSTM with attention
- Input size: 116 sensors × 60 timesteps
- Output size: 35 scenario classes
- Number of LSTM layers: 2
- Loss function: Cross-entropy
- Optimizer: Adam

3.3.LRP Attribution (Baseline)

For the baseline explanation method, we employ LRP attribution (specifically the Gradient × Input variant). For a trained model f classifying input window $\mathbf{x}$ as scenario k, the sensor-level relevance is computed as:

$$\boldsymbol{R_d} = \sum_{t=1}^{T} \left| \boldsymbol{x_{t,d}} \cdot \frac{\partial \boldsymbol{f_k}(\mathrm{x})}{\partial \boldsymbol{x_{t,d}}} \right| \quad (4)$$

The sensor-level relevance R_d is obtained by aggregating absolute relevance values across all 60 timesteps. Sensors are then ranked by their normalized relevance scores. To ensure robustness, relevance scores are averaged across all correctly classified windows (1,353 windows for AB_01; 1,393 for EM_01) and across 15 independent model runs.

The output of this method is a ranked list of sensors with associated relevance scores. This provides sensor identification but no behavioral characterization.

3.4.TCAV-Based Concept Explanation

3.4.1.*Concept Definition*

We define a library of seven behavioral concepts that correspond to recognizable sensor signal patterns:

Table 1 List of Concepts

Concept	Description	Operator Interpretation
high	Signal sustained above 75th percentile	"Parameter is elevated"
low	Signal sustained below 25th percentile	"Parameter is depressed"
very_high	Signal above 95th percentile	"Parameter is critically high"
very_low	Signal below 5th percentile	"Parameter is critically low"
oscillating	Signal exhibits periodic fluctuations	"Parameter is unstable"
step_change	Abrupt change in signal level	"Sudden shift detected"
spike	Brief excursion followed by recovery	"Transient detected"

3.4.2.Concept Example Generation

For each sensor-concept pair (e.g., "Steam Generator 1 Inlet Temperature" + "very_high"), positive and negative examples are extracted from the training data:

(i) **Range-based concepts** (high, low, very_high, very_low): Soft membership functions are used. For instance, a "very_high" concept is activated when

sensor values exceed the 95th percentile of baseline operating conditions.

(ii) **Temporal concepts** (oscillating, spike, step_change): Detected algorithmically—oscillation through autocorrelation peak analysis, step changes through abrupt level shifts in signal derivatives, and spikes through z-score outlier detection ($z > 3.0$).

Each concept is represented by 50 positive and 50 negative example windows.

3.4.3.CAV Training

For each concept C, a Concept Activation Vector (CAV) is trained as follows. Positive and negative example windows are passed through the diagnosis model, and their intermediate activations $\mathbf{a} \in \mathbb{R}^d$ are extracted at the attention layer. A logistic regression classifier is trained on these activations:

$$\mathrm{CAV}_C = \underset{\mathrm{w}}{\mathbf{argmin}} \sum_{i} \boldsymbol{\mathcal{L}}\left(\mathrm{w}^{\top}\mathrm{a}_i, \boldsymbol{y_i}\right) + \boldsymbol{\lambda} \parallel \mathrm{w} \parallel^{\mathbf{2}} \quad (5)$$

where $y_i \in \{0,1\}$ indicates concept presence. The learned weight vector $\mathbf{w}$, normalized to unit length, serves as the CAV—a direction in activation space that encodes the concept. Validation accuracy is computed to ensure the concept is linearly separable in the activation space (typically > 0.7).

3.4.4.TCAV Score Computation

The TCAV score quantifies how much a concept influences the model's classification. For each test window $\mathbf{x}$ from target class k, the directional derivative of the class logit f_k with respect to the activation $\mathbf{a}_l$ at layer l is computed and projected onto the CAV direction:

$$\boldsymbol{S_C}(\mathrm{x}) = \boldsymbol{\nabla}_{\mathrm{a}_l}\boldsymbol{f_k}(\mathrm{x}) \cdot \mathrm{CAV}_C \quad (6)$$

A positive value indicates that moving the activations in the concept direction increases the model's confidence in class k. The TCAV score aggregates this across all test windows:

$$\mathrm{TCAV}_C = \frac{|\{x \in X_k : S_C(x) > 0\}|}{|X_k|} \quad (7)$$

A score of 1.0 means the concept universally drives predictions toward the target class; 0.5 indicates no influence. Statistical significance is assessed via permutation testing. The TCAV importance is computed as $I_C = \max(2 \times \mathrm{TCAV}_C - 1,\ 0)$, mapping the score from [0.5,1.0] to [0,1].

3.4.5.PageRank-Weighted Concept Ranking

Raw TCAV importance alone may rank concepts that are individually influential but peripherally connected in the plant's physical structure. To address this, we weight TCAV importance by each concept's structural centrality in a sensor interaction graph. We construct a directed graph $G = (V, E)$ from time-lagged cross-correlations among sensor-concept activations, where edges represent statistically significant temporal dependencies between concepts. The PageRank centrality (Page et. al. 1999) of each concept node is computed as:

$$\mathrm{PR}(v_i) = \frac{1-\alpha}{|V|} + \alpha \sum_{v_j \in \mathcal{N}^-(v_i)} \frac{\mathrm{PR}(v_j) \cdot w_{ji}}{d^+(v_j)} \quad (8)$$

with damping factor $\alpha = 0.85$ and edge weights derived from interaction strengths. PageRank scores are normalized to [0,1]. The final concept ranking uses a combined score:

$$\mathrm{Score}_C = 0.6 \times I_C + 0.4 \times \mathrm{PR}_C \quad (9)$$

This weighting ensures that top-ranked concepts are both influential for the model's prediction (high TCAV importance) and structurally central in the plant's operational dynamics (high PageRank centrality).

3.4.6.Output Format

The final output is a ranked list of concept-level explanations ordered by combined score. For example:

"For scenario AB_01 (FW Pump #1 Failure), the key diagnostic concepts are: 1. Steam Generator 1 Inlet Temperature is **VERY HIGH** (combined: 0.97) 2. Core Outlet Temperature is **LOW** (combined: 0.96) 3. Cladding Surface

Temperature is **OSCILLATING** (combined: 0.94) …"

Each explanation combines a physical parameter with a behavioral pattern, enabling operators to verify the explanation against current plant conditions.

4. Experiments

4.1.*Dataset*

We evaluate the proposed approach on the iPWR simulation dataset developed by the International Atomic Energy Agency (IAEA). The dataset contains 35 operational scenarios for an integral Pressurized Water Reactor (iPWR) design. Each scenario is simulated at 1 Hz sampling rate across 116 process sensors covering reactor power, temperatures, pressures, flow rates, valve positions, and safety system states. The data was segmented into 60-second windows for input to the diagnosis model. 80% of the dataset was used for training and 20% for testing.

4.2.*Scenarios*

We focus on two representative scenarios that present distinct diagnostic challenges:

Table 2 Experiment Scenario Description

Scenario	Description	Category	Characteristics
AB_01	FW Pump #1 Failure	Abnormal	Loss of forced circulation in primary loop; cascading thermal effects; safety system actuation
EM_01	Loss of Coolant Accident (LOCA)	Emergency	Reactor coolant boundary breach; rapid depressurization; emergency core cooling activation

These scenarios were selected because they represent fundamentally different failure modes.

4.3.*Results: LRP*

Table 3 presents the top-10 sensors identified by LRP attribution for each scenario.

Table 3 Top10 Contributors

R	AB_01	Rel.	EM_01	Rel.
1	Pump speed 2 (rpm)	1.00	Boron concentration (ppm)	1.00
2	Pump speed 1 (rpm)	0.50	Flow to SG 1 (kg/s)	0.92
3	% SG 2 valve opening (%)	0.27	Flow to SG 2 (kg/s)	0.88
4	% SG 1 valve opening (%)	0.26	% Open valve turbine (%)	0.85
5	Condenser inlet temp. (°C)	0.25	Steam flow rate line 1 (kg/s)	0.84
6	DHR1 Pool temp. (°C)	0.19	Steam flow rate line 2 (kg/s)	0.79
7	Containment pressure (bar)	0.18	Steam flow to condenser (kg/s)	0.56
8	Control rod	0.16	Condenser inlet	0.52

R	AB_01	Rel.	EM_01	Rel.
	worth (pcm)		temp. (°C)	
9	Containment temp. (°C)	0.15	Heater 3 outlet temp. (°C)	0.51
10	Pool temp. (°C)	0.14	SG 1 inlet temp. (°C)	0.44

The following observations can be made from the LRP results:

(i) **Direct indicator bias**: For AB_01, the top-ranked sensors are pump speeds—the variables most directly affected by a pump failure. Therefore overlapped information is generated from the LRP.

(ii) **Lack of behavioral context**: The relevance values convey relative importance but provide no information about how each sensor behaves. "Pump speed 2 has relevance 1.0" does not indicate whether the pump is decelerating, oscillating, or has stopped entirely.

(iii) **Relevance distribution**: For AB_01, relevance drops sharply after the top-2 sensors (from 1.0 to 0.27), indicating that the method concentrates attribution on the most obvious indicators. For EM_01, the distribution is more gradual, but the same interpretive limitations apply.

4.4.*Results: TCAV-Based Concept Explanation*

Table 4 presents the top-10 concepts identified by TCAV analysis for each scenario.

Table 4 Concept-Based Reasoning

r	AB_01	Score	EM_01	Score
1	SG 1 Inlet Temp → VERY HIGH	0.97	Relief valve flow → LOW	0.90
2	Core Outlet Temp → LOW	0.96	SG 1 Inlet Temp → LOW	0.89
3	Clad Surface Temp → OSCILLATING	0.94	Power RCP1 → STEP CHANGE	0.87
4	Relief Valve Flow → STEP CHANGE	0.94	PIS Flow 2 → VERY LOW	0.87
5	Rod Reactivity → OSCILLATING	0.91	Steam Flow Condenser → VERY LOW	0.87
6	SG 2 Inlet Temp → LOW	0.90	Core Reactivity → VERY HIGH	0.86
7	Boron Concentration → VERY HIGH	0.89	Heater Power → VERY HIGH	0.86
8	Header FW Pressure → OSCILLATING	0.89	Nuclear Power → HIGH	0.86
9	Coolant Flow Rate → LOW	0.88	DHR2 Pool Temp → HIGH	0.85

r	AB_01	Score	EM_01	Score
10	Coolant Avg Temp → LOW	0.88	Boron Reactivity → HIGH	0.85

The "Score" column represents the combined score (0.6 × TCAV importance + 0.4 × PageRank centrality). All listed concepts have TCAV importance > 0.91, indicating strong model sensitivity.

The TCAV results exhibit the following characteristics:

(i) **Cascading effect identification**: For AB_01, rather than identifying pump speed (the direct failure point), TCAV identifies the downstream thermal consequences: steam generator inlet temperature being "very high" indicates inadequate heat removal, core outlet temperature being "low" reflects the thermal transient propagation, and cladding surface temperature "oscillating" suggests thermal instability at the fuel-cladding interface.

(ii) **Behavioral pattern specificity**: Each concept explanation includes both the physical parameter and its behavioral pattern. "Relief Valve Flow → STEP CHANGE" (rank 4) communicates that the relief valve exhibits sudden actuation, which is the information needed to confirm that the automatic depressurization system has responded to the transient.

(iii) **System-level diagnostic coherence**: For EM_01, the TCAV concepts form a coherent description of a loss-of-coolant event: relief valve flow is low (flow path through the breach), RCP power shows a step change (pump trip on low pressure), safety injection flow is very low (emergency cooling system activation), and steam flow to the condenser is very low (turbine trip and steam line isolation).

5. Discussion

5.1. *Implications for Operator Decision Support*

The experimental results demonstrate that TCAV and LRP provide qualitatively different types of information. LRP identifies which sensors the model relies on most, which is useful for understanding model behavior and for feature selection purposes (Chae et. al. 2025). TCAV identifies what behavioral patterns drive the model's classification, which is more relevant for operator decision support.

In NPP operations, operators are trained to monitor specific behavioral indicators during abnormal situations. Emergency operating procedures (EOPs) describe plant conditions in terms of behavioral states—temperatures rising, pressures decreasing, flow rates oscillating. TCAV explanations align with this format by explicitly stating the behavioral pattern of each identified concept. This alignment means that operators can directly compare the AI's explanation against their training knowledge and current observations without additional interpretation steps.

5.2. *Complementary Use of Methods*

The minimal overlap between LRP and TCAV (5–10%) suggests that these methods are not redundant but rather complementary. LRP is effective for identifying the most sensitive input features—useful for feature selection, dimensionality reduction, and understanding model sensitivity. TCAV is effective for providing behavioral context—useful for operator decision support and reasoning verification. A practical diagnostic system could employ both methods for different purposes.

6. Conclusion

This study presented a TCAV-based operator support system for NPP fault diagnosis and compared it with conventional LRP attribution. The experimental results on the iPWR simulator dataset demonstrate the following:

(i) LRP and TCAV capture fundamentally different aspects of model behavior, with

only 5–10% overlap in identified features across both evaluation scenarios.

(ii) LRP identifies directly affected sensors (e.g., pump speed during a pump failure), while TCAV identifies system-level behavioral patterns (e.g., steam generator inlet temperature being very high).

(iii) TCAV explanations combine physical parameters with behavioral descriptions, providing information that operators can directly verify against their observations and training knowledge.

(iv) The PageRank-weighted concept ranking effectively prioritizes concepts that are both model-influential and structurally central in plant dynamics.

These results suggest that concept-level explanations provide more informative reasoning support for operator decision-making compared to sensor-level attributions. While LRP remains useful for feature selection and model analysis, TCAV-based explanations offer a more suitable format for operational decision support. The proposed approach is applicable to any domain where trained operators need to verify and reason about AI-generated diagnostic information.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (No. RS-2022-00144150)

References

Bae, Junyong, Jong Woo Park, and Seung Jun Lee. 2022. "Limit Surface/States Searching Algorithm with a Deep Neural Network and Monte Carlo Dropout for Nuclear Power Plant Safety Assessment." *Applied Soft Computing* 124 (July): 109007. https://doi.org/10.1016/j.asoc.2022.109007.

Chae, Young Ho, Seung Geun Kim, Jeonghun Choi, Seo Ryong Koo, and Jonghyun Kim. 2025. "Enhancing Nuclear Power Plant Diagnostics: A Comparative Analysis of XAI-Based Feature Selection Methods for Abnormal and Emergency Scenario Detection." *Progress in Nuclear Energy* 185 (July): 105759. https://doi.org/10.1016/j.pnucene.2025.105759.

Kim, Been, Martin Wattenberg, Justin Gilmer, et al. 2018. "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)." arXiv:1711.11279. Preprint, arXiv, June 7. https://doi.org/10.48550/arXiv.1711.11279.

Kim, Taesung, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. 2022. *REVERSIBLE INSTANCE NORMALIZATION FOR ACCURATE TIME-SERIES FORECASTING AGAINST DISTRIBUTION SHIFT*.

Page, Lawrence and Brin, Sergey and Motwani, Rajeev and Winograd, Terry (1999) *The PageRank Citation Ranking: Bringing Order to the Web*. Technical Report. Stanford InfoLab.