Novelty Assessment Report

Paper: Structural Inference: Interpreting Small Language Models with Susceptibilities
PDF URL: https://openreview.net/pdf?id=J4GYMiE3JT
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

We develop a linear response framework for interpretability that treats a neural network as a Bayesian statistical mechanical system. A small perturbation of the data distribution, for example shifting the Pile toward GitHub or legal text, induces a first-order change in the posterior expectation of an observable localized on a chosen component of the network. The resulting susceptibility can be estimated efficiently with local SGLD samples and factorizes into signed, per-token contributions that serve as attribution scores. We combine these susceptibilities into a response matrix whose low-rank structure separates functional modules such as multigram and induction heads in a 3M-parameter transformer.

Core Task Landscape

This paper addresses: **Interpreting Neural Network Components Through Linear Response to Data Distribution Perturbations**
A total of **17 papers** were analyzed and organized into a taxonomy with **11 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Linear Response Theory and Statistical Physics Frameworks**
- **Geometric and Representational Properties of Network Layers**
- **Training Dynamics and Optimization Behavior**
- **Domain-Specific Applications and Physically Interpretable Models**
- **Advanced Architectural and Theoretical Extensions**

Complete Taxonomy Tree

- Interpreting Neural Network Components Through Linear Response to Data Distribution Perturbations Survey Taxonomy
- Linear Response Theory and Statistical Physics Frameworks
  - Bayesian and Statistical Mechanical Interpretability Methods ★ (1 papers)
  - [0] Structural Inference: Interpreting Small Language Models with Susceptibilities (Anon et al., 2026) View paper
  - Neuronal Network Linear Response Models (2 papers)
  - [4] Linear response in neuronal networks: from neurons dynamics to collective response (Bruno Cessac, 2019) View paper
  - [8] How linear response shaped models of neural circuits and the quest for alternatives (T. Herfurth, 2017) View paper
- Geometric and Representational Properties of Network Layers
  - Linear Separability and Piecewise Linear Structure Analysis (3 papers)
  - [1] Understanding Deep Neural Networks via Linear Separability of Hidden Layers (Zhang Chao, 2023) View paper
  - [2] Mean-field Analysis of Piecewise Linear Solutions for Wide ReLU Networks (Alexander P. Shevchenko, 2021) View paper
  - [10] The Evolution of the Interplay Between Input Distributions and Linear Regions in Networks (Qi Xuan, 2023) View paper
  - Representational Drift and Temporal Evolution (1 papers)
  - [12] Stochastic Gradient Descent-induced drift of representation in a two-layer neural network (Pashakhanloo, 2023) View paper
- Training Dynamics and Optimization Behavior
  - Early-Time Learning and Linear Approximations (1 papers)
  - [14] The Surprising Simplicity of the Early-Time Learning Dynamics of Neural Networks (Hu Wei, 2022) View paper
  - Gradient Noise and Stability Analysis (1 papers)
  - [15] On Linear Stability of SGD and Input-Smoothness of Neural Networks (Chao Ma, 2021) View paper
- Domain-Specific Applications and Physically Interpretable Models
  - Physical Systems Modeling with Interpretable Architectures (2 papers)
  - [5] White-box machine learning for uncovering physically interpretable dimensionless governing equations for granular materials (Han Xu, 2025) View paper
  - [7] Physically Interpretable Emulation of a Moist Convecting Atmosphere With a Recurrent Neural Network (Qiyu Song, 2025) View paper
  - Biological Signal Processing and Nonlinear Transformations (3 papers)
  - [6] Nonlinearity from linear scale invariance, quadratic maps and neural network: applications in bio-solitons (Dhurjati Prasad Datta, 2025) View paper
  - [13] A Novel Adaptive Linear Neuron Based on DNA Strand Displacement Reaction Network. (Chengye Zou, 2022) View paper
  - [16] Linear responses to nonlinear signals: A neural network model of spatiotemporal visual processing (Paolo Gaudiano, 1992) View paper
  - Neuronal Dynamics and Chaotic Behavior (1 papers)
  - [3] Chaos in neuronal networks with balanced excitatory and inhibitory activity (C. van Vreeswijk, 1996) View paper

- Advanced Architectural and Theoretical Extensions
  - Polynomial Chaos and Perturbation Theory for Neural Networks (2 papers)
  - [9] The Deep Arbitrary Polynomial Chaos Neural Network or how Deep Artificial Neural Networks could benefit from Data-Driven Homogeneous Chaos Theory (S. Oladyshkin, 2023) View paper
  - [11] Neural network perturbation theory and its application to the Born series (Bastian Kaspschak, 2021) View paper
  - Image Processing with Trainable Diffusion Models (1 papers)
  - [17] Image De-noising with Machine Learning: A (RS Thakur, n.d.) View paper

Narrative

Core task: Interpreting neural network components through linear response to data distribution perturbations. This field seeks to understand how network layers and parameters react when input distributions are slightly perturbed, drawing on concepts from statistical physics and linear response theory. The taxonomy organizes research into several main branches: one rooted in statistical physics frameworks that adapt classical linear response methods to neural systems; another examining geometric and representational properties that emerge in hidden layers; a third focused on training dynamics and how optimization shapes network behavior; a fourth exploring domain-specific applications where physically interpretable models are paramount; and a fifth addressing advanced architectural and theoretical extensions. Representative works span from foundational studies of balanced networks and chaos (Chaos Balanced Networks[3]) to modern analyses of piecewise linear activations (Piecewise Linear ReLU[2]) and separability in hidden representations (Linear Separability Hidden Layers[1]), illustrating how the field bridges classical neuroscience, physics-inspired theory, and contemporary deep learning.

Particularly active lines of work contrast rigorous statistical mechanical approaches—such as those examining neuronal linear response (Linear Response Neuronal[4]) or perturbation expansions (Neural Perturbation Theory[11])—with more applied efforts that embed physical constraints into architectures for domains like granular materials (White-box Granular Materials[5]) or convection modeling (Physically Interpretable Convection[7]). A central tension involves balancing theoretical rigor with practical interpretability: some studies pursue exact characterizations of linear regions and input distributions (Input Distributions Linear Regions[10]), while others prioritize domain-specific fidelity. The original paper, Structural Inference Susceptibilities[0], sits within the Bayesian and statistical mechanical interpretability branch, emphasizing how susceptibility measures—borrowed from physics—can reveal structural properties of network components under distributional shifts. This approach aligns closely with works like Linear Response Neuronal[4] in its physics-inspired framing, yet contrasts with more geometry-focused studies (Linear Separability Hidden Layers[1]) by prioritizing statistical inference over purely representational analysis.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

Both subtopics apply linear response theory and statistical mechanics frameworks to understand network behavior, but target fundamentally different domains. The original leaf focuses on interpreting artificial neural networks (machine learning models) through Bayesian and statistical mechanical lenses to extract attribution scores and susceptibilities. The sibling focuses on biological neuronal networks, deriving explicit linear responses to external stimuli using Gibbs distributions and ergodic theory.

Similarities: - Both employ statistical mechanics and linear response theory as core mathematical frameworks - Both aim to characterize network behavior through perturbation analysis - Both leverage probabilistic/statistical descriptions of network states (Bayesian methods vs. Gibbs distributions)

Differences: - Application domain: artificial neural networks (ML models) vs. biological neuronal networks - Goal: interpretability and attribution in ML vs. understanding stimulus response in neuroscience - Methodological emphasis: Bayesian inference and susceptibilities for ML interpretability vs. ergodic theory and explicit response derivations for neuronal dynamics - Exclusion boundaries explicitly separate ML interpretability from neurobiological modeling

Suggested Search Directions: - Explore whether statistical mechanical interpretability methods from ML could inform analysis of biological neural networks - Investigate transfer of linear response techniques between artificial and biological neural systems - Search for hybrid approaches that bridge interpretability in both artificial and neuronal networks using shared statistical mechanics foundations

Sibling Subtopics

- **Neuronal Network Linear Response Models** (leaves: 1, papers: 2)
- Scope: Derives explicit linear response of neuronal networks to external stimuli using Gibbs distributions and ergodic theory.
- Exclude: Excludes artificial neural network interpretability methods; see Bayesian and Statistical Mechanical Interpretability Methods.

Contributions Analysis

Overall novelty summary. The paper introduces a linear response framework that treats neural networks as Bayesian statistical mechanical systems, deriving susceptibilities to data distribution perturbations and using these to identify functional modules. Within the taxonomy, it occupies the 'Bayesian and Statistical Mechanical Interpretability Methods' leaf under 'Linear Response Theory and Statistical Physics Frameworks'. Notably, this leaf contains only the original paper itself—no sibling papers are present—indicating this represents a relatively sparse and novel research direction within the broader field of neural network interpretability through perturbation analysis.

The taxonomy reveals that neighboring work exists primarily in two directions: neuronal network models applying linear response to biological systems (two papers in 'Neuronal Network Linear Response Models') and geometric analyses of layer representations without explicit statistical mechanical framing (four papers across geometric subtopics). The original paper bridges these areas by applying physics-inspired linear response theory specifically to artificial neural networks for interpretability, rather than modeling biological neurons or analyzing static geometric properties. This positioning suggests the work synthesizes concepts from adjacent branches—statistical physics formalism and interpretability goals—in a combination not extensively explored by prior literature.

Among the 30 candidates examined across three contributions, none were identified as clearly refuting any claimed novelty. The 'Susceptibility framework' contribution examined 10 candidates with zero refutable matches, as did the 'Structural inference methodology' and 'Per-token attribution scores' contributions. This absence of overlapping prior work across all contributions, combined with the limited search scope, suggests that within the examined literature the specific combination of Bayesian statistical mechanics, susceptibility-based attribution, and modular structure inference appears relatively unexplored. However, the search examined only top-30 semantic matches, leaving open the possibility of relevant work outside this scope.

Given the limited search scale and the paper's placement in an otherwise unpopulated taxonomy leaf, the work appears to occupy a genuinely novel intersection of statistical physics and neural network interpretability. The absence of sibling papers and zero refutable candidates across 30 examined works supports this impression, though the analysis cannot rule out relevant prior work beyond the top-K semantic neighborhood. The framework's distinctiveness lies in its integrated approach—combining Bayesian mechanics, local sampling, and low-rank factorization—rather than any single component in isolation.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Susceptibility framework for neural network interpretability

Description: The authors develop a linear response framework rooted in statistical physics and Bayesian learning theory that treats neural networks as statistical mechanical systems. Susceptibilities measure how infinitesimal perturbations to the data distribution induce first-order changes in the expected behavior of network components, providing a principled link between data structure and model internals.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Prediction Accuracy & Reliability: Classification and Object Localization Under Distribution Shift

URL: View paper

Brief Assessment

Prediction Under Shift[46] focuses on prediction accuracy and uncertainty quantification under distribution shift for classification and object localization tasks in autonomous driving. It does not address neural network interpretability through susceptibilities measuring component responses to data distribution changes.

---

2. Ensemble-based deep reinforcement learning for vehicle routing problems under distribution shift

URL: View paper

Brief Assessment

Ensemble Vehicle Routing[44] focuses on ensemble-based deep reinforcement learning for vehicle routing problems under distribution shift, not on interpretability frameworks for neural networks or susceptibility measures for data distribution changes.

---

3. Adaptive State Estimation and Continual Learning under Data Distribution Shift

URL: View paper

Brief Assessment

Adaptive Continual Learning[41] focuses on distribution-shift detection and continual learning for state estimation, not on interpretability frameworks using susceptibilities to measure neural network component responses to data distribution changes.

---

4. Intermediate layer classifiers for ood generalization

URL: View paper

Brief Assessment

Intermediate Layer Classifiers[43] focuses on using intermediate layer representations for out-of-distribution generalization in classification tasks, not on developing a statistical physics framework for measuring neural network component responses to data distribution changes through susceptibilities.

---

5. Dish-ts: a general paradigm for alleviating distribution shift in time series forecasting

URL: View paper

Brief Assessment

Dish-ts Distribution Shift[42] focuses on distribution shift in time series forecasting using coefficient networks, not on neural network interpretability through susceptibilities measuring responses to data distribution perturbations in the context of statistical physics and Bayesian learning theory.

---

6. Measuring domain shift for deep learning in histopathology

URL: View paper

Brief Assessment

Domain Shift Histopathology[40] focuses on measuring domain shift in histopathology using representation statistics, not on developing a susceptibility framework rooted in statistical physics and Bayesian learning theory for neural network interpretability.

---

7. Incomplete Multisource Domain Adaptation for Fault Diagnosis of Blast Furnace

URL: View paper

Brief Assessment

Incomplete Multisource Adaptation[47] focuses on fault diagnosis in blast furnaces using transfer learning and domain adaptation techniques. It does not address neural network interpretability, susceptibilities, or statistical mechanical frameworks for understanding model components.

---

8. Diagnosing model performance under distribution shift

URL: View paper

Brief Assessment

Diagnosing Distribution Shift[38] focuses on decomposing model performance degradation across distribution shifts in deployment settings, not on interpreting neural network internals through susceptibilities to data distribution perturbations as a statistical mechanical system.

---

9. Measuring robustness to natural distribution shifts in image classification

URL: View paper

Brief Assessment

Robustness Natural Shifts[39] focuses on measuring robustness to natural distribution shifts in image classification, not on developing susceptibility frameworks for neural network interpretability or treating networks as statistical mechanical systems.

---

10. Agreement-on-the-line: Predicting the performance of neural networks under distribution shift

URL: View paper

Brief Assessment

Agreement-on-the-line[45] focuses on predicting neural network performance under distribution shift using agreement between model pairs, not on susceptibilities measuring component responses to data distribution changes as a statistical mechanical framework for interpretability.

Contribution 2: Structural inference methodology

Description: The authors present a methodology that uses response matrices of susceptibilities combined with PCA to identify functional modules and internal structure in neural networks. This approach reveals how models balance expression and suppression of patterns, enabling discovery of circuits like induction heads through data-driven analysis.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. AIFS--ECMWF's data-driven forecasting system

URL: View paper

Brief Assessment

AIFS Data-Driven Forecasting[21] focuses on weather forecasting using graph neural networks and transformers, not on discovering internal structure in neural networks through susceptibility analysis or PCA-based functional module identification.

2. A unifying principle for the functional organization of visual cortex

URL: View paper

Brief Assessment

Visual Cortex Organization[23] focuses on spatial organization of visual cortex neurons using topographic constraints in neural networks, not on discovering internal functional modules through data-driven susceptibility analysis of language models.

3. An ai-driven, scalable, and modular digital twin framework for traffic management

URL: View paper

Brief Assessment

Digital Twin Traffic[22] focuses on traffic management using digital twin architecture and deep learning for traffic prediction. It does not address neural network interpretability, internal structure discovery, or susceptibility-based analysis methods.

4. Data-driven discovery of intrinsic dynamics

URL: View paper

Brief Assessment

Intrinsic Dynamics Discovery[26] focuses on discovering low-dimensional manifolds and dynamics from time series data using charts and atlases, not on interpreting neural network internal structure through susceptibility analysis and response matrices.

5. Sensor-fault detection, isolation and accommodation for digital twins via modular data-driven architecture

URL: View paper

Brief Assessment

Sensor-Fault Digital Twins[24] focuses on sensor fault detection in industrial systems using neural network estimators and classifiers, not on discovering internal structure in neural networks through data-driven analysis of susceptibilities or response matrices.

6. FE: an efficient data-driven multiscale approach based on physics-constrained neural networks and automated data mining

URL: View paper

Brief Assessment

Physics-Constrained Multiscale[20] focuses on data-driven multiscale modeling for materials using physics-constrained neural networks and computational homogenization, not on discovering internal structure and functional modules in neural networks through response matrices and PCA.

7. Dynamic system modeling using a multisource transfer learning-based modular neural network for industrial application

URL: View paper

Brief Assessment

Multisource Transfer Learning[25] focuses on transfer learning for modular neural networks in industrial process modeling, not on discovering internal structure through data-driven analysis of susceptibilities or response matrices.

8. Data-Driven Object Tracking: Integrating Modular Neural Networks into a Kalman Framework

URL: View paper

Brief Assessment

Data-Driven Object Tracking[19] focuses on multi-object tracking using neural networks integrated into Kalman filters for autonomous driving applications, not on discovering internal structure in neural networks through data-driven analysis of susceptibilities.

9. Big data and fuzzy logic for demand forecasting in supply chain management: a data-driven approach

URL: View paper

Brief Assessment

Fuzzy Demand Forecasting[18] focuses on supply chain demand forecasting using fuzzy logic and big data analytics, not on discovering internal structure in neural networks through data-driven analysis of susceptibilities or response matrices.

10. Data-driven emergence of convolutional structure in neural networks

URL: View paper

Brief Assessment

Convolutional Structure Emergence[27] focuses on how neural networks autonomously develop convolutional structure from translation-invariant image data, not on discovering functional modules in language models through susceptibility-based analysis of data distribution perturbations.

Contribution 3: Per-token susceptibility attribution scores

Description: The authors show that susceptibilities can be decomposed into per-token contributions with interpretable signs (positive for suppression, negative for expression). These token-level attribution scores can be efficiently estimated using local Stochastic Gradient Langevin Dynamics sampling around network checkpoints.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. GrAInS: Gradient-based Attribution for Inference-Time Steering of LLMs and VLMs

URL: View paper

Brief Assessment

GrAInS Gradient Attribution[29] focuses on token-level attribution for inference-time steering in LLMs/VLMs using integrated gradients, while the original paper develops per-token susceptibility scores within a Bayesian statistical mechanical framework for interpretability. These are fundamentally different methodological approaches serving different purposes.

2. Transformer-based protein generation with regularized latent space optimization

URL: View paper

Brief Assessment

Protein Generation Latent[32] focuses on protein generation with latent space optimization and token-level representations for proteins, not on per-token attribution scores for neural network interpretability estimated via stochastic gradient sampling.

3. Bayesian Influence Functions for Scalable Data Attribution

URL: View paper

Brief Assessment

Bayesian Influence Functions[36] focuses on training data attribution via covariance-based influence functions, not on per-token attribution scores for interpretability. The methods and objectives differ fundamentally from the original paper's susceptibility framework.

4. Interferometric token resonance: A computational framework for measuring intramodel signal interference in large language models

URL: View paper

Brief Assessment

Interferometric Token Resonance[28] focuses on measuring intramodel signal interference in LLMs, not on per-token attribution scores estimated via stochastic gradient sampling for Bayesian interpretability frameworks.

5. On identifiability in transformers

URL: View paper

Brief Assessment

Identifiability Transformers[33] focuses on attention weight identifiability and gradient-based attribution for understanding token mixing in transformers, not on per-token attribution scores estimated via stochastic gradient sampling for Bayesian neural networks as described in the original paper.

6. Evaluating attribution methods for explainable nlp with transformers

URL: View paper

Brief Assessment

Attribution Methods Transformers[35] focuses on evaluating existing attribution methods for NLP transformers (gradients, attention), not on developing susceptibility-based attribution scores estimated via SGLD sampling.

7. Unmasking Backdoors: An Explainable Defense via Gradient-Attention Anomaly Scoring for Pre-trained Language Models

URL: View paper

Brief Assessment

Gradient-Attention Anomaly Scoring[30] focuses on backdoor defense in NLP models using attention and gradient signals for anomaly detection, not on per-token attribution scores for general neural networks estimated with stochastic gradient sampling as in the original paper's susceptibility framework.

8. LibraGrad: Balancing Gradient Flow for Universally Better Vision Transformer Attributions

URL: View paper

Brief Assessment

LibraGrad Balancing Flow[34] addresses gradient flow imbalances in vision transformers for attribution methods, not per-token attribution scores for language models estimated via stochastic gradient sampling. The technical domains and methodologies are fundamentally different.

9. Probabilistic contextual resonance in large language model decoding through selfmodulated semantic interference

URL: View paper

Brief Assessment

Contextual Resonance LLM[31] focuses on semantic interference and resonance penalties in LLM decoding, not on per-token attribution scores estimated via stochastic gradient sampling for interpretability purposes.

10. Where Did It Go Wrong? Attributing Undesirable LLM Behaviors via Representation Gradient Tracing

URL: View paper

Brief Assessment

Representation Gradient Tracing[37] focuses on token-level attribution via representation gradients for diagnosing undesirable LLM behaviors, not on decomposing susceptibilities into per-token contributions using SGLD sampling around network checkpoints as in the original paper.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Structural Inference: Interpreting Small Language Models with Susceptibilities View paper
- [1] Understanding Deep Neural Networks via Linear Separability of Hidden Layers View paper
- [2] Mean-field Analysis of Piecewise Linear Solutions for Wide ReLU Networks View paper

- [3] Chaos in neuronal networks with balanced excitatory and inhibitory activity View paper
- [4] Linear response in neuronal networks: from neurons dynamics to collective response View paper
- [5] White-box machine learning for uncovering physically interpretable dimensionless governing equations for granular materials View paper
- [6] Nonlinearity from linear scale invariance, quadratic maps and neural network: applications in bio-solitons View paper
- [7] Physically Interpretable Emulation of a Moist Convecting Atmosphere With a Recurrent Neural Network View paper
- [8] How linear response shaped models of neural circuits and the quest for alternatives View paper
- [9] The Deep Arbitrary Polynomial Chaos Neural Network or how Deep Artificial Neural Networks could benefit from Data-Driven Homogeneous Chaos Theory View paper
- [10] The Evolution of the Interplay Between Input Distributions and Linear Regions in Networks View paper
- [11] Neural network perturbation theory and its application to the Born series View paper
- [12] Stochastic Gradient Descent-induced drift of representation in a two-layer neural network View paper
- [13] A Novel Adaptive Linear Neuron Based on DNA Strand Displacement Reaction Network. View paper
- [14] The Surprising Simplicity of the Early-Time Learning Dynamics of Neural Networks View paper
- [15] On Linear Stability of SGD and Input-Smoothness of Neural Networks View paper
- [16] Linear responses to nonlinear signals: A neural network model of spatiotemporal visual processing View paper
- [17] Image De-noising with Machine Learning: A View paper
- [18] Big data and fuzzy logic for demand forecasting in supply chain management: a data-driven approach View paper
- [19] Data-Driven Object Tracking: Integrating Modular Neural Networks into a Kalman Framework View paper
- [20] FE: an efficient data-driven multiscale approach based on physics-constrained neural networks and automated data mining View paper
- [21] AIFS--ECMWF's data-driven forecasting system View paper
- [22] An ai-driven, scalable, and modular digital twin framework for traffic management View paper
- [23] A unifying principle for the functional organization of visual cortex View paper
- [24] Sensor-fault detection, isolation and accommodation for digital twins via modular data-driven architecture View paper
- [25] Dynamic system modeling using a multisource transfer learning-based modular neural network for industrial application View paper
- [26] Data-driven discovery of intrinsic dynamics View paper
- [27] Data-driven emergence of convolutional structure in neural networks View paper
- [28] Interferometric token resonance: A computational framework for measuring intramodel signal interference in large language models View paper
- [29] GrAInS: Gradient-based Attribution for Inference-Time Steering of LLMs and VLMs View paper
- [30] Unmasking Backdoors: An Explainable Defense via Gradient-Attention Anomaly Scoring for Pre-trained Language Models View paper
- [31] Probabilistic contextual resonance in large language model decoding through selfmodulated semantic interference View paper
- [32] Transformer-based protein generation with regularized latent space optimization View paper
- [33] On identifiability in transformers View paper
- [34] LibraGrad: Balancing Gradient Flow for Universally Better Vision Transformer Attributions View paper
- [35] Evaluating attribution methods for explainable nlp with transformers View paper
- [36] Bayesian Influence Functions for Scalable Data Attribution View paper
- [37] Where Did It Go Wrong? Attributing Undesirable LLM Behaviors via Representation Gradient Tracing View paper
- [38] Diagnosing model performance under distribution shift View paper
- [39] Measuring robustness to natural distribution shifts in image classification View paper
- [40] Measuring domain shift for deep learning in histopathology View paper
- [41] Adaptive State Estimation and Continual Learning under Data Distribution Shift View paper
- [42] Dish-ts: a general paradigm for alleviating distribution shift in time series forecasting View paper
- [43] Intermediate layer classifiers for ood generalization View paper
- [44] Ensemble-based deep reinforcement learning for vehicle routing problems under distribution shift View paper
- [45] Agreement-on-the-line: Predicting the performance of neural networks under distribution shift View paper
- [46] Prediction Accuracy & Reliability: Classification and Object Localization Under Distribution Shift View paper
- [47] Incomplete Multisource Domain Adaptation for Fault Diagnosis of Blast Furnace View paper