# Novelty Assessment Report

**Paper**: Provably Explaining Neural Additive Models
**PDF URL**: https://openreview.net/pdf?id=040ClRXMf3
**Venue**: ICLR 2026 Conference Submission
**Year**: 2026
**Report Generated**: 2026-01-05

## Abstract

Despite significant progress in post-hoc explanation methods for neural networks, many remain heuristic and lack provable guarantees. A key approach for obtaining explanations with provable guarantees is by identifying a cardinally-minimal subset of input features which by itself is provably sufficient to determine the prediction. However, for standard neural networks, this task is often computationally infeasible, as it demands a worst-case exponential number of verification queries in the number of input features, each of which is NP-hard. In this work, we show that for Neural Additive Models (NAMs), a recent and more interpretable neural network family, we can efficiently generate explanations with such guarantees. We present a new model-specific algorithm for NAMs that generates provably cardinally-minimal explanations using only a logarithmic number of verification queries in the number of input features, after a parallelized preprocessing step with logarithmic runtime in the required precision is applied to each small univariate NAM component. Our algorithm not only makes the task of obtaining cardinally-minimal explanations feasible, but even outperforms existing algorithms designed to find subset-minimal explanations -- which may be larger and less informative but easier to compute -- despite our algorithm solving a much more difficult task. Our experiments demonstrate that, compared to previous algorithms, our approach provides provably smaller explanations than existing works and substantially reduces the computation time. Moreover, we show that our generated provable explanations offer benefits that are unattainable by standard sampling-based techniques typically used to interpret NAMs.



## Core Task Landscape

This paper addresses: **Generating Provably Cardinally-Minimal Sufficient Explanations for Neural Additive Models**

A total of **2 papers** were analyzed and organized into a taxonomy with **3 categories**.

### Taxonomy Overview

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

- **Provable Explanation Generation for Neural Additive Models**
- **Interpretable Model Applications Using Additive Structures**

### Complete Taxonomy Tree

- Generating Provably Cardinally-Minimal Sufficient Explanations for Neural Additive Models Survey Taxonomy
- Provable Explanation Generation for Neural Additive Models
  - Cardinally-Minimal Sufficient Explanations with Verification ★ (1 papers)
  - [0] Provably Explaining Neural Additive Models (Anon et al., 2026) View paper
- Interpretable Model Applications Using Additive Structures
  - Neural Additive Models for Clustering (1 papers)
  - [1] NeurCAM: Interpretable Neural Clustering via Additive Models (Cohen Eldan, 2024) View paper
  - Additive Models for Multi-Criteria Decision Aiding (1 papers)
  - [2] Necessary and Sufficient Explanations of Multi-Criteria Decision Aiding Models, with and Without Interacting Criteria (Johanne Cohen, 2023) View paper

### Narrative

Core task: Generating provably cardinally-minimal sufficient explanations for neural additive models. The field centers on making neural additive models—architectures that decompose predictions into interpretable per-feature contributions—more transparent through rigorous explanation methods. The taxonomy divides into two main branches: one focused on provable explanation generation, which emphasizes formal guarantees and verification that explanations are both minimal and sufficient, and another on interpretable model applications that leverage additive structures for practical deployment. The provable branch tends to concentrate on algorithmic techniques that certify explanation quality, ensuring that no smaller subset of features can justify a prediction, while the applications branch explores how additive decompositions can be used in domains requiring inherent interpretability.

Within the provable explanation generation branch, recent work has explored different notions of minimality and sufficiency. Some studies develop verification procedures to confirm that a chosen feature subset is indeed cardinally minimal, while others investigate necessary and sufficient conditions for explanations, as seen in Necessary Sufficient Explanations[2]. NeurCAM[1] exemplifies efforts to integrate class activation mapping ideas with neural additive architectures, bridging visualization and formal explanation. The original paper, Explaining Neural Additive[0], sits squarely in the provable branch, emphasizing cardinally-minimal sufficient explanations with verification. Compared to neighboring works, it appears to prioritize formal guarantees of minimality over heuristic or approximate methods, aiming to certify that no redundant features remain in the explanation while maintaining prediction sufficiency.

## Related Works in Same Category

No sibling papers and no sibling subtopics were found under the same parent taxonomy node; the paper appears structurally isolated in the taxonomy.

## Contributions Analysis

**Overall novelty summary.** The paper develops a model-specific algorithm for generating cardinally-minimal sufficient explanations in Neural Additive Models (NAMs), reducing verification complexity from exponential to logarithmic in the number of input features. Within the taxonomy, it occupies the sole position in the 'Cardinally-Minimal Sufficient Explanations with Verification' leaf under 'Provable Explanation Generation for Neural Additive Models'. This leaf contains only the original paper itself, indicating a sparse research direction with no sibling papers identified in the taxonomy structure.

The taxonomy reveals two main branches: provable explanation generation (where this work resides) and interpretable model applications using additive structures. Neighboring work includes Neural Additive Models for Clustering and Additive Models for Multi-Criteria Decision Aiding, both focused on practical applications rather than formal guarantees. The taxonomy narrative mentions related efforts like NeurCAM and Necessary Sufficient Explanations, which explore verification procedures and different notions of minimality, suggesting the paper connects to a broader interest in certified explanations but diverges by targeting cardinality optimality specifically for NAMs.

Among 11 candidates examined across three contributions, no refutable prior work was identified. The 'First provably sufficient explanations for NAMs' contribution examined 10 candidates with none providing overlapping prior work, while the 'Parallel interval importance sorting procedure' examined 1 candidate without refutation. The 'Model-specific algorithm' contribution examined no candidates. Given the limited search scope of 11 papers total, these statistics suggest the specific combination of cardinality minimality, provable sufficiency, and NAM-specific algorithms may be relatively unexplored, though the analysis does not cover exhaustive literature.

Based on the limited search scope and sparse taxonomy position, the work appears to address a gap in providing formal guarantees for NAM explanations. However, the analysis covers only top-K semantic matches and does not exhaustively survey all explanation methods for additive models or verification techniques in interpretable ML. The absence of sibling papers and limited candidate examination suggest either genuine novelty in this specific problem formulation or incomplete coverage of related verification-based explanation work.

---

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

### Contribution 1: Model-specific algorithm for cardinally-minimal explanations in NAMs

**Description**: The authors introduce an algorithm tailored to Neural Additive Models that efficiently computes cardinally-minimal sufficient explanations. Unlike general neural networks requiring exponential queries, this method exploits NAMs' additive structure to achieve logarithmic query complexity through parallelized preprocessing and binary search.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

### Contribution 2: Parallel interval importance sorting procedure

**Description**: The authors develop a preprocessing stage that operates in parallel on each univariate NAM component to compute importance intervals and establish a total ordering of features. This parallelized approach substantially reduces computational overhead by working on small univariate functions rather than the full model.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

#### 1. Feature selection for classification of SELDI-TOF-MS proteomic profiles

**URL**: View paper

**Brief Assessment**

SELDI Feature Selection[12] focuses on feature selection for proteomic mass spectrometry profiles using univariate statistical methods and correlation-based filtering, not on parallel preprocessing for neural additive models with importance intervals.

---

### Contribution 3: First provably sufficient explanations for NAMs

**Description**: The authors present the first approach for generating explanations with provable sufficiency guarantees specifically for Neural Additive Models. This advances the trustworthiness of NAMs in safety-critical applications where formal guarantees are essential.

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. On the role of context in reading time prediction

**URL**: View paper

**Brief Assessment**

Context Reading Time[6] focuses on reading time prediction using language models and contextual predictors (surprisal, PMI). It does not address provably sufficient explanations for Neural Additive Models or formal verification guarantees.

---

#### 2. Predictive model and risk analysis for peripheral vascular disease in type 2 diabetes mellitus patients using machine learning and shapley additive explanation

**URL**: View paper

**Brief Assessment**

Diabetes Vascular SHAP[4] applies SHAP to medical prediction tasks for peripheral vascular disease, not to Neural Additive Models with formal sufficiency guarantees. The technical domains and objectives are entirely different.

---

#### 3. Contextual Importance and Utility: aTheoretical Foundation

**URL**: View paper

**Brief Assessment**

Contextual Importance Utility[8] focuses on contextual importance and utility theory for model-agnostic explanations, not on provably sufficient explanations with formal guarantees for Neural Additive Models. The theoretical foundations and technical approaches are fundamentally different.

---

#### 4. Unifying Attribution-Based Explanations Using Functional Decomposition

**URL**: View paper

**Brief Assessment**

Unifying Attribution Decomposition[9] focuses on unifying attribution-based explanation methods through game-theoretic frameworks and additive decompositions, not on provably sufficient explanations with formal guarantees for Neural Additive Models.

---

### 5. Step-wise explanations for the additive model

**URL**: View paper

**Brief Assessment**

Stepwise Additive Explanations[3] focuses on additive models in decision-making contexts (comparing alternatives based on criteria), not neural additive models (NAMs) for machine learning predictions with formal verification guarantees.

---

### 6. FAIXID: A framework for enhancing AI explainability of intrusion detection results using data cleaning techniques

**URL**: View paper

**Brief Assessment**

FAIXID[7] focuses on intrusion detection systems using data cleaning techniques for AI explainability. It does not address provably sufficient explanations with formal guarantees for Neural Additive Models, which is the core novelty claim of the original paper.

---

### 7. Weakly-Supervised Abstraction for Linear Additive Models

**URL**: View paper

**Brief Assessment**

Weakly Supervised Abstraction[5] focuses on causal abstraction for linear additive models in a different context (aggregating low-level causal variables while preserving interventional distributions), not on generating provably sufficient explanations with formal guarantees for Neural Additive Models as interpretability tools.

---

### 8. A Meta-Learning Control Algorithm with Provable Finite-Time Guarantees

**URL**: View paper

**Brief Assessment**

Meta Learning Control[11] focuses on online meta-learning control algorithms for linear deterministic systems with regret guarantees, not on explainability methods or Neural Additive Models. The technical domains are entirely different.

---

### 9. A Layered Approach to Safety Certification for AI-Driven Systems Using Explainable and Verifiable Machine Learning Models

**URL**: View paper

**Brief Assessment**

Layered Safety Certification[10] focuses on safety certification frameworks for AI systems broadly, not on provably sufficient explanations for Neural Additive Models specifically. The candidate discusses formal verification and explainability in certification contexts but does not address cardinally-minimal explanations for NAMs.

---

### 10. NeurCAM: Interpretable Neural Clustering via Additive Models

**URL**: View paper

**Brief Assessment**

NeurCAM[1] focuses on interpretable clustering using neural additive models with fuzzy membership and additive explanations. It does not address provably sufficient explanations with formal guarantees for NAMs, which is the core novelty claim of the original paper.

---

## Appendix: Text Similarity Detection

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

## References


- [0] Provably Explaining Neural Additive Models View paper
- [1] NeurCAM: Interpretable Neural Clustering via Additive Models View paper
- [2] Necessary and Sufficient Explanations of Multi-Criteria Decision Aiding Models, with and Without Interacting Criteria View paper
- [3] Step-wise explanations for the additive model View paper
- [4] Predictive model and risk analysis for peripheral vascular disease in type 2 diabetes mellitus patients using machine learning and shapley additive explanation View paper
- [5] Weakly-Supervised Abstraction for Linear Additive Models View paper
- [6] On the role of context in reading time prediction View paper
- [7] FAIXID: A framework for enhancing AI explainability of intrusion detection results using data cleaning techniques View paper
- [8] Contextual Importance and Utility: aTheoretical Foundation View paper
- [9] Unifying Attribution-Based Explanations Using Functional Decomposition View paper
- [10] A Layered Approach to Safety Certification for AI-Driven Systems Using Explainable and Verifiable Machine Learning Models View paper
- [11] A Meta-Learning Control Algorithm with Provable Finite-Time Guarantees View paper
- [12] Feature selection for classification of SELDI-TOF-MS proteomic profiles View paper