PROVABLY EXPLAINING NEURAL ADDITIVE MODELS

Anonymous authors
Paper under double-blind review

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 *(globally) cardinal-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 (globally) cardinal-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 (globally) cardinal minimal explanations feasible, but even outperforms existing algorithms designed to find *(locally) 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.

1 INTRODUCTION

Various methods have been proposed to explain neural network predictions. Classic additive feature attribution approaches – such as LIME (Ribeiro et al., 2016), SHAP (Lundberg & Lee, 2017), and IG (Sundararajan et al., 2017) – assume near-linear behavior in a local region around the instance. Other methods, like Anchors (Ribeiro et al., 2018) and SIS (Carter et al., 2019), aim to identify a (nearly) sufficient subset of input features – referred to here as an *explanation* – that determines the prediction. While Anchors and SIS rely on probabilistic sampling and lack provable sufficiency guarantees, recent work has shown that neural network verification tools can serve as a backbone for generating provably sufficient explanations (Wu et al., 2023; Bassan & Katz, 2023; La Malfa et al., 2021; Izza et al., 2024), making them particularly valuable in safety-critical domains (Marques-Silva & Ignatiev, 2022). In this context, *smaller* sufficient explanations are typically preferred, as *minimality* is considered an additional key interpretability property (Ignatiev et al., 2019; Carter et al., 2019; Darwiche & Hirth, 2020; Ribeiro et al., 2018; Barceló et al., 2020b).

However, while such explanations are highly desirable, generating them for standard neural networks is notoriously computationally challenging (Barceló et al., 2020b). In particular, obtaining *(globally) cardinal-minimal* explanations, requires, in the worst case, an *exponential* number of neural network verification queries (Barceló et al., 2020b; Ignatiev et al., 2019; Bassan & Katz, 2023), each being NP-hard (Katz et al., 2017; Sälzer & Lange, 2021), rendering the task infeasible even for toy examples (Ignatiev et al., 2019). Consequently, existing methods focus on *(locally) subset-minimal* explanations (Wu et al., 2023; Bassan & Katz, 2023; Bassan et al., 2025a), which are typically suboptimal in size, potentially large, and thus less informative than their globally minimal counterparts. Moreover, even these approaches remain limited to relatively small models, as they still require a linear number of verification queries (Wu et al., 2023; Bassan et al., 2025b).

Our contributions. Since computing (globally) cardinal-minimal sufficient explanations is provably intractable for general neural networks, a natural question arises: *Can certain neural architectures with more interpretable structures enable efficient computation of such explanations?* Although this task remains challenging even for simplified models such as *binarized* neural networks or those with a *single* hidden layer (Adolfi et al., 2025; Barceló et al., 2020a; Sälzer & Lange, 2021), we show in this work that it becomes tractable for a different class of neural architectures previously unexplored in this context: *Neural Additive Models (NAMs)* (Agarwal et al., 2021). NAMs are a widely adopted architecture that has received significant attention in recent years (Agarwal et al., 2021; Radenovic et al., 2022; Bechler-Speicher et al., 2024; Kim et al., 2024; Zhang et al., 2024). By enforcing an additive structure over input features, NAMs support interpretable, per-feature contributions while maintaining the expressive capabilities of neural networks.

Our NAM-specific algorithm vs. previous algorithms. Unlike existing algorithms that require a linear number of verification queries in the number of input features for locally minimal explanations – or an exponential number for globally minimal ones – our approach exploits the additive structure of NAMs to compute provably *globally* minimal explanation subsets using only a *logarithmic* number of verification queries. This is realized by introducing a highly parallelized preprocessing step – each operating independently on a small univariate component of the NAM – enabling a substantial overall efficiency gain. As a result, our method yields explanations that are both provably sufficient and cardinal-minimal, far more efficiently than standard algorithms typically applied to neural networks. Experiments using state-of-the-art verifiers confirm that our algorithm generates explanations substantially faster and produces notably smaller subsets than prior approaches.

Our provable NAM explanations vs. standard sampling interpretations. NAMs are typically interpreted by sampling input points and visualizing the behavior of each univariate function (Agarwal et al., 2021; Radenovic et al., 2022). Thanks to their additive structure, these models allow per-feature contributions to be examined individually. We show that purely sampling-based methods can yield misleading interpretations, whereas our provably sufficient explanations avoid this by design, underscoring their importance in safety-critical domains.

Overall, our work advances explanations with provable guarantees in two ways: (i) it introduces the first method for certifiable explanations in NAMs, boosting their trustworthiness in safety-critical settings, and (ii) by *efficiently* generating provable explanations, NAMs – unlike general neural networks – open a path toward interpretable architectures where such guarantees can be derived at scale. A central challenge ahead is to design models that balance high expressivity and accuracy with efficient provable explanations, and we view our work as a significant first step in that direction.

2 PRELIMINARIES

2.1 NOTATION

We denote scalars with lower-case letters, vectors with bold lower-case letters, and sets in calligraphic font. The i-th entry of a vector $\mathbf{x}$ is denoted by $\mathbf{x}_{(i)}$. For $n \in \mathbb{N}$, let $[n] := \{1, \ldots, n\}$.

2.2 NEURAL NETWORK VERIFICATION

Neural network verification aims to verify certain input-output relationships of neural networks. For a neural network $f\colon \mathbb{R}^n \to \mathbb{R}^c$, a neural network verifier *formally proves* that there does not exist an input $\mathbf{x} \in \mathbb{R}^n$ where both an input specification $\psi_{\text{in}}(\mathbf{x})$, and an *unsafe* output specification $\psi_{\text{out}}(f(\mathbf{x}))$ hold at the same time. Although this problem is NP-hard (Katz et al., 2017; Sälzer & Lange, 2021), these tools have seen rapid scalability improvements in recent years (Brix et al., 2024).

2.3 NEURAL ADDITIVE MODELS (NAMS).

A neural additive model (NAM) f for a *regression* task, where $f : \mathbb{R}^n \to \mathbb{R}$, is defined as:

$$f(\mathbf{x}) := \beta_0 + \sum_{i=1}^{k} f_i(\mathbf{x}_{(i)}), \tag{1}$$

where each $f_i : \mathbb{R} \to \mathbb{R}$ is a univariate neural network, $\beta_0 \in \mathbb{R}$ is the intercept, and f denotes the full NAM. For *binary classification*, we assume that an additional step function is applied: $f(\mathbf{x}) := \text{step}(\beta_0 + \sum_{i=1}^{k} f_i(\mathbf{x}_{(i)}))$, where $\text{step}(z) = 1$ if $z \geq 0$ and $\text{step}(z) = 0$ otherwise. In the *multi-class* setting with c classes, the logit for class $j \in [c]$ is given by $f_j(\mathbf{x}) := \beta_{j,0} + \sum_{i=1}^{k} f_{j,i}(\mathbf{x}_{(i)})$, and the model predicts $f(\mathbf{x}) := \arg\max_{j \in [c]} f_j(\mathbf{x})$.

In this work, we develop algorithms for all three settings – (i) regression, (ii) binary classification, and (iii) multi-class classification – but for clarity, we focus the main presentation on the binary classification case, with extensions for the other settings provided in Appendix C.

3 Provably Sufficient Explanations for Neural Networks

We begin by reviewing standard algorithms developed for *general* neural networks that identify provably locally or globally minimal sufficient explanations. We note that we focus on *post-hoc* sufficient explanations for a specific input $\mathbf{x} \in \mathbb{R}^n$, i.e., for the output $f(\mathbf{x})$, and post-hoc indicates that the explanation is generated after the model has been trained.

Sufficient Explanations. A common method for interpreting the decisions of classifiers involves identifying subsets of input features $\mathcal{S} \subseteq [n]$ such that fixing these features to their specific values guarantees the prediction remains unchanged. Specifically, these techniques guarantee that the classification result remains consistent across *any* potential assignment within the complementary set $\bar{\mathcal{S}} := [n] \setminus \mathcal{S}$. While in the classic setting features in the complementary set $\bar{\mathcal{S}}$ are allowed to take on any possible feature values (Ignatiev et al., 2019; Darwiche & Hirth, 2020; Bassan & Katz, 2023), a more feasible and generalizable version restricts the possible assignments for $\bar{\mathcal{S}}$ to a bounded ϵ_p-region (Wu et al., 2023; La Malfa et al., 2021; Izza et al., 2024). We use $(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) \in \mathbb{R}^n$ to denote an assignment where the features of $\mathcal{S}$ are set to the values of the vector $\mathbf{x} \in \mathbb{R}^n$ and the features of $\bar{\mathcal{S}}$ are set to the values of another vector $\tilde{\mathbf{x}} \in \mathbb{R}^n$ within the ϵ_p-region.

Definition 1 (Sufficient Explanation)**.** *Given a neural network f, an input $\boldsymbol{x} \in \mathbb{R}^n$, a perturbation radius $\epsilon_p \in \mathbb{R}_+$, and a subset $\mathcal{S} \subseteq [n]$, we say that $\mathcal{S}$ is a sufficient explanation concerning the query $\langle f, \boldsymbol{x}, \mathcal{S}, \epsilon_p \rangle$ on an ℓ_p-norm ball $B_p^{\epsilon_p}$ of radius $\epsilon_p \in \mathbb{R}_+$ around $\boldsymbol{x}$ iff it holds that:*

$$\forall \tilde{\boldsymbol{x}} \in B_p^{\epsilon_p}(\boldsymbol{x}) \colon \quad f(\boldsymbol{x}_{\mathcal{S}}; \tilde{\boldsymbol{x}}_{\bar{\mathcal{S}}}) = f(\boldsymbol{x}), \qquad \text{with } B_p^{\epsilon_p}(\boldsymbol{x}) := \{\tilde{\boldsymbol{x}} \in \mathbb{R}^n \mid \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_p \leq \epsilon_p\}.$$

We define $\text{suff}(f, \boldsymbol{x}, \mathcal{S}, \epsilon_p) = 1$ *iff $\mathcal{S}$ constitutes a sufficient explanation with respect to the query $\langle f, \boldsymbol{x}, \mathcal{S}, \epsilon_p \rangle$, and* $\text{suff}(f, \boldsymbol{x}, \mathcal{S}, \epsilon_p) = 0$ *otherwise.*

Def. 1 can be formulated as a neural network verification query (Sec. 2.2). This method has been proposed by prior studies, which employed these techniques to validate the sufficiency of specific subsets (Wu et al., 2023; Bassan & Katz, 2023; La Malfa et al., 2021; Izza et al., 2024).

Minimal Explanations. Evidently, selecting the entire input set as the subset $\mathcal{S}$, that is, setting $\mathcal{S} := [n]$, yields a sufficient explanation. Nonetheless, the prevailing consensus in the literature is that smaller subsets tend to be more informative or meaningful (Ribeiro et al., 2018; Carter et al., 2019; Barceló et al., 2020b; Ignatiev et al., 2019). Consequently, there is considerable interest in identifying subsets that are not only sufficient but also satisfy some notion of minimality. We focus on two specific minimality criteria: (global) cardinality minimality and (local) subset minimality.

Definition 2 (Minimal Sufficient Explanations)**.** *Given a neural network f, an input $\boldsymbol{x} \in \mathbb{R}^n$, and a subset $\mathcal{S} \subseteq [n]$ that is a sufficient explanation concerning $\langle f, \boldsymbol{x}, \mathcal{S}, \epsilon_p \rangle$ on $B_p^{\epsilon_p}$ of radius ϵ_p, then:*

1. *We say that $\mathcal{S}$ is a (globally)* cardinal-minimal *sufficient explanation (Barceló et al., 2020a; Bassan et al., 2024) concerning $\langle f, \boldsymbol{x}, \mathcal{S}, \epsilon_p \rangle$ iff there does not exist a sufficient explanation $\mathcal{S}'$ concerning $\langle f, \boldsymbol{x}, \mathcal{S}', \epsilon_p \rangle$ with $|\mathcal{S}'| < |\mathcal{S}|$).*
2. *We say that $\mathcal{S}$ is a (locally)* subset-minimal *sufficient explanation (Arenas et al., 2022; Ignatiev et al., 2019) concerning $\langle f, \boldsymbol{x}, \mathcal{S}, \epsilon_p \rangle$ iff any $\mathcal{S}' \subset \mathcal{S}$ is not a sufficient explanation concerning $\langle f, \boldsymbol{x}, \mathcal{S}', \epsilon_p \rangle$.*

Minimal sufficient explanations can also be determined using neural network verifiers. This process requires executing multiple verification queries to ensure the minimality of the subset. Alg. 1 outlines such a procedure (Ignatiev et al., 2019; Wu et al., 2023; Bassan & Katz, 2023). The algorithm

begins with an explanation $\mathcal{S}$ encompassing the entire feature set $[n]$ and iteratively tries to exclude a feature i from $\mathcal{S}$, each time checking whether $\mathcal{S} \setminus \{i\}$ remains sufficient. If $\mathcal{S} \setminus \{i\}$ is still sufficient, feature i is removed; otherwise, it is retained in the explanation. This process is repeated until a subset-minimal sufficient explanation is obtained.

Algorithm 1 Greedy Subset Minimal Explanation Search

Input: Neural network $f\colon \mathbb{R}^n \to \mathbb{R}^c$, input $\mathbf{x} \in \mathbb{R}^n$, perturbation radius $\epsilon_p \in \mathbb{R}_+$

$\mathcal{S} \leftarrow [n]$
for each feature $i \in [n]$ **do** ▷ $\text{suff}(f, \mathbf{x}, \mathcal{S}, \epsilon_p)$ holds
  if $\text{suff}(f, \mathbf{x}, \mathcal{S} \setminus \{i\}, \epsilon_p)$ **then**
    $\mathcal{S} \leftarrow \mathcal{S} \setminus \{i\}$
  end if
end for
return $\mathcal{S}$ ▷ $\mathcal{S}$ is a *subset-minimal* explanation concerning $\langle f, \mathbf{x}, \mathcal{S}, \epsilon_p \rangle$

4 Provably Sufficient Explanations for NAMs

While generating (globally) cardinal-minimal sufficient explanations is computationally expensive for general neural networks, we present a highly efficient algorithm for NAMs in this section. NAMs are generally considered very interpretable due to the univariate functions, but sufficient guarantees can only be obtained through formal verification to avoid misleading conclusions (Fig. 1). Our algorithm consists of two main stages: (i) As a preprocessing step, we compute an "importance" interval for each feature $i \in [n]$ based on the univariate functions f_i to obtain a total ordering. (ii) This allows us to perform a binary search over the sorted intervals to identify the cardinal-minimal sufficient explanation.

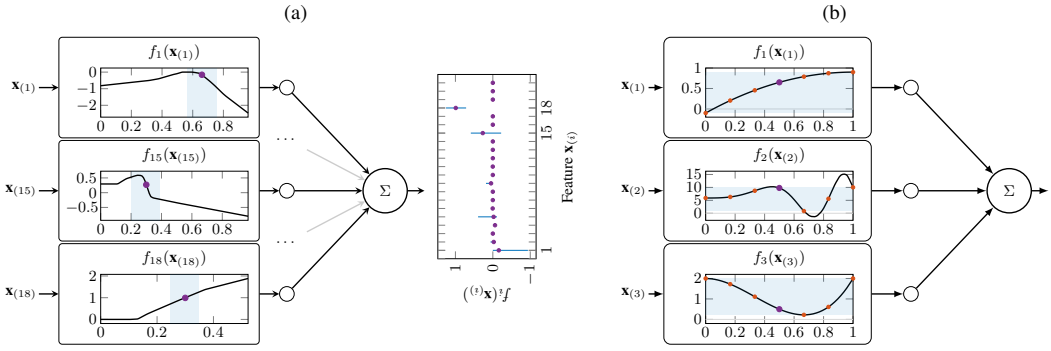

Figure 1: Sufficient explanations in NAMs: (a) Users must examine the neighborhood of an input for proper interpretation; e.g, users might wrongly conclude that feature 1 from the FICO HELOC dataset alone determines a positive output, but small changes in features 15 or 18 can flip the classification. (b) Outputs of continuous neighborhoods can be misleading if not verified, since sampling may miss extrema; e.g., users might wrongly believe that only feature 1 yields negative outputs, while feature 2 can also flip the classification.

4.1 Stage 1 — Parallel Interval Importance Sorting

This subsection outlines the first stage of our algorithm, which involves determining an ordering of the feature importance to be used in the subsequent phase. We define feature $i \in [n]$ as more "important" than feature $j \neq i$ if perturbing feature i leads to a greater deviation in the final prediction $f(\mathbf{x})$ compared to perturbing feature j. In particular, we measure the derivation towards the decision boundary to flip the classification. Thanks to the additive structure of the NAM, this analysis can be conducted independently for each univariate component $f_i\colon \mathbb{R} \to \mathbb{R}$, allowing for a direct comparison of their individual importance.

Without loss of generality, let us assume that our binary classifier predicts $f(\mathbf{x}) = 1$, meaning that $\beta_0 + \sum_{i=1}^{k} f_i(\mathbf{x}_{(i)}) \geq 0$ (the case $f(\mathbf{x}) = 0$ follows symmetrically). In this setting, we perturb the input to each univariate component f_i individually and measure how much the overall prediction decreases, i.e., towards the decision boundary. Therefore, for each $i \in [n]$, we want to find

$$\mathbf{x}^*_{(i)} = \underset{\tilde{\mathbf{x}}_{(i)} \in \mathcal{B}_p^{\epsilon_p}(\mathbf{x}_{(i)})}{\arg\min} \; f_i(\tilde{\mathbf{x}}_{(i)}). \tag{2}$$

Accurately determining these minimal values is usually computationally infeasible Katz et al. (2017). Fortunately, we do not need to find the exact minimum but only bounds $[l_i, u_i] \subset \mathbb{R}$ such that $l_i \leq f_i(x^*_{(i)}) \leq u_i$, up to a precision such that a total order over the input features can be obtained. The procedure is outlined in Alg. 2.

Algorithm 2 Parallel Interval Importance Sorting

Input: NAM f, input $\mathbf{x} \in \mathbb{R}^n$, perturbation radius $\epsilon_p \in \mathbb{R}_+$

for each feature $i \in [n]$ *in parallel* **do**
  Extract initial bounds α_i, β_i for $f_i(\tilde{\mathbf{x}}_i)$ such that $\tilde{\mathbf{x}}_{(i)} \in \mathcal{B}_p^{\epsilon_p}(\mathbf{x}_{(i)})$
  $l_i \leftarrow \alpha_i$, $u_i \leftarrow \beta_i$
  while True **do**
    $m_i \leftarrow \frac{l_i + u_i}{2}$
    if verify($\forall \tilde{\mathbf{x}}_i \in \mathcal{B}_p^{\epsilon_p}(\mathbf{x}_{(i)}),\ f_i(\tilde{\mathbf{x}}_{(i)}) \geq m_i$) **then**
      $l_i \leftarrow m_i$
    else
      $u_i \leftarrow m_i$
    end if
    $\Delta l_i \leftarrow f_i(\mathbf{x}_{(i)}) - l_i$; $\Delta u_i \leftarrow f_i(\mathbf{x}_{(i)}) - u_i$
    if for all $i \neq j$ it holds that: $\Delta u_i \geq \Delta l_j$ or $\Delta u_j \geq \Delta l_i$ **then**
      break
    end if
  end while
end for
return $\arg\text{sort}([(\Delta l_1, \Delta u_1), \ldots, (\Delta l_n, \Delta u_n)])$ in ascending order

Alg. 2 operates in parallel across each univariate component of the NAM. For each component, the algorithm begins by issuing an incomplete verification query to obtain an initial lower bound $f_i(\tilde{\mathbf{x}}_{(i)})$, denoted by l_i and u_i, evaluated over the domain $\tilde{\mathbf{x}}_{(i)} \in \mathcal{B}_p^{\epsilon_p}(\mathbf{x}_{(i)})$. Subsequently, each thread independently conducts a binary search using verification queries to iteratively refine $[l_i, u_i]$, narrowing in on the true lower bound $f_i(x^*_{(i)})$. To quantify the deviation of the unperturbed output $f_i(\mathbf{x}_{(i)})$ from these bounds, we define the relative differences Δl_i and Δu_i. After each iteration within every parallel thread, the algorithm evaluates whether a full, non-overlapping sorting of all pairs $(\Delta l_1, \Delta u_1), \ldots, (\Delta l_n, \Delta u_n)$ is possible. If such an ordering cannot yet be achieved, the bounds get iteratively refined. Finally, the ordering according to the determined importance is returned.

While the initially computed bounds α_i, β_i enclose the entire output domain of each component $i \in [n]$, the binary search narrows the bounds down around the minimum $f_i(\mathbf{x}^*_{(i)})$ (or maximum for $f(\mathbf{x}) = 0$); thus, no longer covering the entire domain. As $f_i(\mathbf{x}^*_{(i)})$ is just a scalar, this total order can be determined using a complete verifier.

This non-overlapping ordering provides a rigorous measure for the impact of perturbing a single feature through its corresponding component function $f_i(\mathbf{x}_{(i)})$. This sorting forms the foundation for the next phase of the algorithm, which identifies a provably cardinal-minimal sufficient subset. The following proposition formalizes this first step of the argument:

Proposition 1. *Given a NAM f, an input $\boldsymbol{x} \in \mathbb{R}^n$ and a perturbation radius $\epsilon_p \in \mathbb{R}_+$, let Alg. 2 return a total list order over the input features according to their importance. Then, the following holds: For any sufficient explanation $\mathcal{S}$ that includes feature i, and for any feature $j \notin \mathcal{S}$ such that $i \prec j$ in the list ordering, the set $\mathcal{S} \setminus \{i\} \cup \{j\}$ is also a sufficient explanation.*

All proofs are provided in Appendix A. Intuitively, this proposition shows that for any two features i and j, the ordering produced by Alg. 2 defines a notion of "importance" such that if feature i appears in a sufficient explanation, and feature j does not, then feature i can always be replaced by j. This implies that, in this context, j is at least as "important" as i. This property is crucial for later leveraging the extracted ordering to construct provably cardinal-minimal explanations.

Complexity. The complexity of Alg. 2 is governed by the use of ρ parallel processors, where each processor independently carries out a binary search. This binary search iteratively partitions based on the Δu_i and Δl_i bounds and terminates once the bounds of two distinct univariate components no longer overlap. Given the initial gap between the upper and lower bounds, $\alpha_i - f_i(\mathbf{x}_{(i)})$, for each component f_i, and the precision for component f_i defined by its minimal separation from the adjacent features in the sorted ordering – namely, $\xi_i := \min\{|\hat{\Delta}l_{i+1} - \hat{\Delta}u_i|, |\hat{\Delta}l_i - \hat{\Delta}u_{i-1}|\}$, with $\hat{\Delta}l_\square, \hat{\Delta}u_\square$ denoting the bounds in the last iteration. We can now prove that the number of neural network verifier calls is bounded by a (parallelized) *logarithmic* term, as formalized in the following proposition. Limitations and optimizations are further discussed in Appendix B.

Proposition 2. *Given ρ parallelized processors, Alg. 2, performs an overall number of $T_\rho(n) = \mathcal{O}\Big(\big(\frac{n}{\rho}\big)\log\big(\max_{i\in[n]}\big(\frac{\beta_i-\alpha_i}{\xi_i}\big)\big)\Big) \xrightarrow[\rho\to n]{} \mathcal{O}\Big(\log\big(\max_{i\in[n]}\big(\frac{\beta_i-\alpha_i}{\xi_i}\big)\big)\Big)$ calls to the neural network verifier, each on a $f_i(\cdot)$ component, where $\xi_i := \min\{|\hat{\Delta}l_{i+1} - \hat{\Delta}u_i|, |\hat{\Delta}l_i - \hat{\Delta}u_{i-1}|\}$.*

4.2 Stage 2 — Feature Selection Based on the Derived Feature Intervals

In this subsection, we will describe the second part of our algorithm that can obtain a provably cardinal-minimal sufficient explanation, given the derived interval orderings that were obtained from Alg. 2. To the total order, we can apply a binary search to obtain the explanation, resulting in a logarithmic number of verification queries in the number of input features. However, to simplify the presentation of this algorithm, we will start by presenting a naive greedy approach that runs in a linear number of steps, and then move on to presenting the binary-search approach. The naive approach is depicted in Alg. 3.

Algorithm 3 Greedy Cardinal-Minimal Linear Explanation Search

Input: NAM f, input $\mathbf{x} \in \mathbb{R}^n$, perturbation radius $\epsilon_p \in \mathbb{R}_+$

```
S ← [n]
for each feature i ∈ [n], ordered by Alg. 2 do        ▷ suff(f, x, S, ε_p) holds
    if suff(f, x, S \ {i}, ε_p) then
        S ← S \ {i}
    end if
end for
return S                                              ▷ S is a cardinal-minimal explanation concerning ⟨f, x, S, ε_p⟩
```

Alg. 3 closely mirrors the operation of Alg. 1: It begins by initializing the explanation $\mathcal{S}$ to the full feature set $[n]$, and then iteratively removes features, updating $\mathcal{S} \leftarrow \mathcal{S} \setminus \{i\}$, until reaching a minimal explanation. However, unlike Alg. 1, which is only guaranteed to converge to a (locally) *subset-minimal* explanation, Alg. 3 is designed to converge to the more challenging objective of finding a (globally) *cardinal-minimal* sufficient explanation. This stronger guarantee is enabled by the total ordering $(\hat{\Delta}l_i, \hat{\Delta}ui)_{i=1}^n$ computed by Alg. 2, which ranks the features by their importance. This leads to the following proposition:

Proposition 3. *Given a NAM f, an input $\boldsymbol{x} \in \mathbb{R}^n$, and a perturbation radius $\epsilon_p \in \mathbb{R}_+$, Alg. 3 performs $\mathcal{O}(n)$ queries and returns a* cardinal-minimal *sufficient explanation. This stands in contrast to Alg. 1, which is only guaranteed to return a* subset-minimal *sufficient explanation.*

Alg. 3 can be significantly enhanced by replacing the linear ordering with a binary search strategy (Alg. 4). Crucially, this step is not possible with the naive, unsorted approach (Alg. 1), as it does not guarantee convergence to a cardinal-minimal explanation, and may not even yield a *subset-minimal* explanation. This is because, in an arbitrary feature ordering, there may be multiple points at which a non-sufficient subset becomes sufficient, making the binary search unreliable. In contrast, the preprocessing step in Alg. 2 imposes a structured sorting of features, which allows Alg. 4 to reliably converge to a *cardinal-minimal* sufficient explanation using only a *logarithmic* number of queries.

Algorithm 4 Greedy Cardinal-Minimal Logarithmic Explanation Search

Input: NAM f, input $\mathbf{x} \in \mathbb{R}^n$, perturbation radius $\epsilon_p \in \mathbb{R}_+$

$F \leftarrow$ total order of features (Alg. 2)
$l \leftarrow 1$; $u \leftarrow n$
while $l \neq u$ **do**
  $m \leftarrow \lfloor \frac{l+u}{2} \rfloor$
  if $\text{suff}(f, \mathbf{x}, \{F[1], \dots F[m]\}, \epsilon_p)$ **then**
    $l \leftarrow m$
  else
    $u \leftarrow m - 1$
  end if
end while
$\mathcal{S} \leftarrow \{F[1], \dots F[m]\}$
return $\mathcal{S}$ ▷ $\mathcal{S}$ is a *cardinal-minimal* explanation concerning $\langle f, \mathbf{x}, \mathcal{S}, \epsilon_p \rangle$

Proposition 4. *Given a NAM f, an input $\boldsymbol{x} \in \mathbb{R}^n$, and a perturbation radius $\epsilon_p \in \mathbb{R}_+$, Alg. 4 performs $\mathcal{O}(\log(n))$ queries and returns a* cardinal-minimal *sufficient explanation.*

Overall complexity results. By combining Alg. 2 with Alg. 4, we obtain a cardinal-minimal explanation for $\langle f, \mathbf{x}, \epsilon_p \rangle$. This unified algorithm yields a substantial efficiency gain, reducing the worst-case requirement of an *exponential* number of verification queries to only a *logarithmic* number of (parallelized) queries. The first segment of these queries operate by running verification queries on *univariate components* f_i of the model, which are far smaller, and hence more efficient to verify than direct queries to f. The resulting complexity bound is formalized in the following theorem:

Theorem 1. *Running Alg. 2 and Alg. 4 obtains a* cardinal-minimal *sufficient explanation with $\mathcal{O}\Big((\frac{n}{\rho})\log(\max_{i\in[n]}\big(\frac{\beta_i-\alpha_i}{\xi_i}\big)\Big) \xrightarrow[\rho\to n]{} \mathcal{O}\Big(\log\big(\max_{i\in[n]}\big(\frac{\beta_i-\alpha_i}{\xi_i}\big)\Big)$ queries to $f_i(\cdot)$ components, plus $\mathcal{O}(\log n)$ queries to $f(\cdot)$. In contrast, standard algorithms require $\mathcal{O}(2^n)$ verification queries to $f(\cdot)$ for a cardinal-minimal explanation, or $\mathcal{O}(n)$ verification queries to $f(\cdot)$ for only a subset-minimal explanation.*

5 Evaluation

Experimental Setup. We implemented our main algorithmic approach (Alg. 2 followed by Alg. 4) using α-β-CROWN as the backend verifier, the current state-of-the-art in neural network verification (Wang et al., 2021; Zhou et al., 2024; Kotha et al., 2023; Brix et al., 2024; Chiu et al., 2025). We conducted extensive experiments on four widely used tabular-data benchmarks in the context of NAMs (Agarwal et al., 2021; Radenovic et al., 2022): (i) Breast Cancer, (ii) CREDIT, (iii) FICO HELOC, all of which are prominent in safety-critical domains. We adopted the same model architectures as prior work in the NAM literature (Agarwal et al., 2021; Radenovic et al., 2022). Evaluation details, additional experiments, and ablation studies are in Appendix D.

5.1 Our Algorithm vs. Previous Algorithms

We begin by comparing our results with prior algorithms proposed in the literature for obtaining provably minimal sufficient explanations. Since our method targets the much stronger notion of (globally) cardinal-minimal sufficient explanations for the first time, any naive baseline – that computes such explanations by exhaustively enumerating all 2^n input subsets, verifies their sufficiency, and selects the one with the smallest cardinality – would not finish with reasonable timeouts. Thus, we compare our approach to the more scalable task of finding (locally) subset-minimal explanations, a weaker notion of minimality, using the standard greedy algorithm employed by previous works (Wu et al., 2023; Bassan et al., 2025a; Izza et al., 2024; Ignatiev et al., 2019; La Malfa et al., 2021) (Alg. 1). Because subset-minimal explanations depend on feature orderings, we consider two setups: (i) a basic lexicographic ordering of features, and (ii) a more sophisticated reverse-sensitivity ordering, following prior approaches (Wu et al., 2023; Bassan et al., 2025a; Izza et al., 2024; Wu et al., 2024).

Table 1: Comparison of average explanation size and computation time.

	Breast Cancer		CREDIT		FICO HELOC	
Method	Size (↓)	Time [s] (↓)	Size (↓)	Time [s] (↓)	Size (↓)	Time [s] (↓)
Ours	**4.00±4.24**	**35.60±1.34**	**3.76**±2.62	**132.67±36.76**	**5.59**±1.80	317.92±222.07
Lexicographic	16.58±5.44	634.92±77.23	12.42±6.45	473.63±128.38	15.60±7.53	**146.16**±188.07
Sensitivity	16.27±5.57	636.79±87.44	3.82±1.84	407.93±126.63	9.45±5.90	250.09±148.44

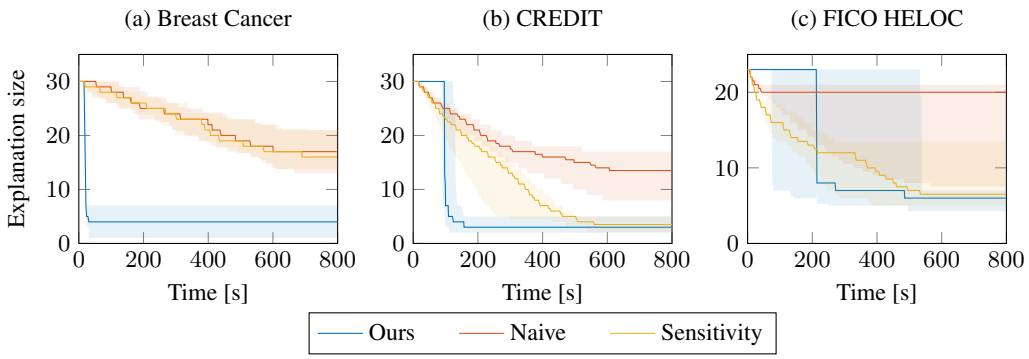

Figure 2: Explanation size over time for all datasets.

The results in Tab. 1 demonstrate that our proposed algorithm achieves substantial improvements in both computation time and explanation size over previous algorithms. Beyond reducing explanation size compared to standard subset-minimal explanation algorithms (which follows by necessity, since we enforce a stronger notion of minimality), our method also achieves substantial runtime gains, despite solving a much harder task. This advantage stems from our NAM-specific algorithm, which requires only a *logarithmic* number of parallelized queries – rather than a linear number – executed over univariate components f_i, which are much faster to verify.

5.2 Explanation Progression in Time

To further assess the advantages of our algorithm over prior methods, we analyze the explanation sizes produced by our approach in comparison to subset-minimal methods and track their evolution over time. This analysis is illustrated in Fig. 2. The results show that while subset-minimal approaches converge slowly and often stagnate in local minima. Our method – though it begins later due to the preprocessing step in Alg. 2, which sorts features with only a logarithmic number of parallelized queries – quickly outpaces them once sorting is complete. At this point, it requires significantly fewer queries, relying only on a binary search over the sorted features as in Alg. 4. This second phase is not only substantially faster but also provably attains the cardinal-minimal explanation, i.e., the global optimum, unlike subset-minimal approaches.

5.3 Comparison to Purely Sampling-Based Methods

NAMs are generally viewed as very interpretable as their univariate functions f_i for each feature allow for simple visualizations (such as in Fig. 1). Most commonly, these visualizations are obtained through sampling over the respective feature domain to get a good approximation of each univariate function. However, we show in this experiment that this discretization through sampling and the resulting interpretations can be misleading. Peaks and other extrema that are missed through sampling can lead to insufficient explanations, which can be fatal in safety-critical domains. An extreme case is abstractly depicted in Fig. 1b, but we have also observed insufficient explanations in practice. To demonstrate this, we evaluate $1,000$ evenly-spaced samples instead of each verification query. After the explanations are generated, we test their sufficiency using α, β-CROWN (Tab. 2): On the CREDIT and FICO HELOC datasets, more than half of the explanations obtained through sampling could not be verified. In contrast, all our explanations are sufficient by construction.

Table 2: Comparison against a purely sampling-based approach.

	CREDIT			FICO HELOC		
Method	Size (↓)	Time [s] (↓)	Sufficiency [%] (↑)	Size (↓)	Time [s] (↓)	Sufficiency [%] (↑)
Ours	3.76±2.62	132.67±36.76	**100.00**	5.59±1.80	317.92±222.07	**100.00**
Sampling	**2.67**±3.10	**5.10±0.26**	31.37	**3.04**±3.05	**6.89±2.57**	25.49

6 Related Work

Formal XAI. Our work relates to the field of *formal XAI* (Marques-Silva, 2023), which seeks explanations with provable guarantees. Prior efforts have developed sufficient explanations for models such as decision trees (Huang et al., 2021; Bounia & Koriche, 2023), linear models (Marques-Silva et al., 2020; Subercaseaux et al., 2025), monotonic classifiers (Marques-Silva et al., 2021), and tree ensembles (Izza & Marques-Silva, 2021; Ignatiev et al., 2022; Audemard et al., 2022; 2023). Closer to our setting are works on minimal sufficient explanations for neural networks (La Malfa et al., 2021; Wu et al., 2023; Izza et al., 2024; Bassan et al., 2025a), which rely on neural network verification queries. While such verifiers have become more scalable in recent years, computing such explanations is still costly, often requiring many (linear or exponential) verification queries (Ignatiev et al., 2019). Our method takes a first step toward reducing this cost by focusing on neural network families with interpretable structure, and in particular on NAMs.

Neural Additive Models (NAMs). NAMs extend *Generalized Additive Models (GAMs)* (Hastie, 2017; Nelder & Wedderburn, 1972), a classic interpretable family of ML models (Caruana et al., 2015; Zhong et al., 2023; Liu et al., 2022; Bordt & von Luxburg, 2023; Enouen & Liu, 2025; Chen et al., 2020), by replacing each univariate component with a neural network, thereby combining interpretability with expressivity. First introduced by (Agarwal et al., 2021), NAMs achieved competitive accuracy on tabular tasks and were applied in healthcare and COVID-19 modeling. Subsequent works suggested potential refinements of their training and architecture (Radenovic et al., 2022; Chang et al., 2021; Bouchiat et al., 2024; Xu et al., 2023) and proposed additional variants (Bechler-Speicher et al., 2024; Jiao et al., 2024) and applications (Thielmann et al., 2024).

7 Limitations

Like all methods that obtain provably minimal and sufficient explanations, our approach depends on invoking neural network verification queries, which do not yet scale to state-of-the-art models. Still, neural network verification has advanced rapidly in recent years (Brix et al., 2024), and the scalability of our approach will improve alongside it. Importantly, our method offers two substantially critical improvements: (i) it reduces the number of queries from exponential (or linear, in relaxed tasks) to *logarithmic*, and (ii) it operates on *univariate* components f_i, where verification is far cheaper since the certified models are small and more interpretable by design compared to the entire large model f. Together, these make our algorithm far more practical for NAMs, and we show that it indeed efficiently produces explanations on standard benchmarks where prior algorithms fail.

8 Conclusion

Provably minimal and sufficient explanations represent a highly desirable goal in explainability, as they offer certifiable guarantees on both faithfulness and conciseness. For standard neural networks, however, this task is computationally prohibitive, requiring an exponential number of verification queries. We present a NAM-specific algorithm that reduces the complexity from *exponential* to *logarithmic* parallelized queries, achieving dramatic gains in both speed and explanation size. Moreover, we show that these explanations reveal insights into NAMs that sampling-based methods cannot capture. Our work thus makes provable explanations feasible in practice and opens the door to extending them across other interpretable neural network families.

REFERENCES

Federico Adolfi, Martina Vilas, and Todd Wareham. The Computational Complexity of Circuit Discovery for Inner Interpretability. In *Proc. 13th Int. Conf. on Learning Representations (ICLR)*, 2025.

Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey E Hinton. Neural Additive Models: Interpretable Machine Learning with Neural Nets. *Advances in neural information processing systems (NeurIPS)*, 34:4699–4711, 2021.

Marcelo Arenas, Pablo Barceló, Miguel Romero Orth, and Bernardo Subercaseaux. On Computing Probabilistic Explanations for Decision Trees. In *Proc. 35th Int. Conf. on the Advances in Neural Information Processing Systems (NeurIPS)*, pp. 28695–28707, 2022.

Gilles Audemard, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis. Trading Complexity for Sparsity in Random Forest Explanations. In *Proc. 36th AAAI Conf. on Artificial Intelligence*, pp. 5461–5469, 2022.

Gilles Audemard, Jean-Marie Lagniez, Pierre Marquis, and Nicolas Szczepanski. Computing Abductive Explanations for Boosted Trees. In *Proc. Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, pp. 4699–4711, 2023.

P. Barceló, M. Monet, J. Pérez, and B. Subercaseaux. Model interpretability through the lens of computational complexity. *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 15487–15498, 2020a.

Pablo Barceló, Mikaël Monet, Jorge Pérez, and Bernardo Subercaseaux. Model Interpretability Through the Lens of Computational Complexity. *Proc. 33rd Int. Conf. on Advances in Neural Information Processing Systems (NeurIPS)*, pp. 15487–15498, 2020b.

Shahaf Bassan and Guy Katz. Towards Formal XAI: Formally Approximate Minimal Explanations of Neural Networks. In *Proc. 29th Int. Conf. on Tools and Algorithms for the Construction and Analysis of Systems (TACAS)*, pp. 187–207, 2023.

Shahaf Bassan, Guy Amir, and Guy Katz. Local vs. Global Interpretability: A Computational Complexity Perspective. In *Proc. 41st Int. Conf. on Machine Learning (ICML)*, pp. 3133–3167, 2024.

Shahaf Bassan, Yizhak Yisrael Elboher, Tobias Ladner, Matthias Althoff, and Guy Katz. Explaining, Fast and Slow: Abstraction and Refinement of Provable Explanations. In *Forty-second International Conference on Machine Learning (ICML)*, 2025a.

Shahaf Bassan, Ron Eliav, and Shlomit Gur. Explain Yourself, Briefly! Self-Explaining Neural Networks with Concise Sufficient Reasons. In *Proc. 13th Int. Conf. on Learning Representations (ICLR)*, 2025b.

Maya Bechler-Speicher, Amir Globerson, and Ran Gilad-Bachrach. The Intelligible and Effective Graph Neural Additive Network. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:90552–90578, 2024.

Sebastian Bordt and Ulrike von Luxburg. From Shapley Values to Generalized Additive Models and Back. In *International Conference on Artificial Intelligence and Statistics*, pp. 709–745, 2023.

Kouroche Bouchiat, Alexander Immer, Hugo Yèche, Gunnar Ratsch, and Vincent Fortuin. Improving neural additive models with bayesian principles. In *International Conference on Machine Learning (ICML)*, pp. 4416–4443, 2024.

Louenas Bounia and Frederic Koriche. Approximating Probabilistic Explanations via Supermodular Minimization. In *Proc. 39th Int. Conf. on Uncertainty in Artificial Intelligence (UAI)*, pp. 216–225, 2023.

Christopher Brix, Stanley Bak, Taylor T Johnson, and Haoze Wu. The fifth international verification of neural networks competition (VNN-COMP 2024): Summary and results. *arXiv preprint arXiv:2412.19985*, 2024.

Brandon Carter, Jonas Mueller, Siddhartha Jain, and David Gifford. What Made You Do This? Understanding Black-Box Decisions with Sufficient Input Subsets. In *Proc. 22nd Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, pp. 567–576, 2019.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1721–1730, 2015.

Chun-Hao Chang, Rich Caruana, and Anna Goldenberg. Node-gam: Neural generalized additive model for interpretable deep learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Hong Chen, Yingjie Wang, Feng Zheng, Cheng Deng, and Heng Huang. Sparse Modal Additive Model. *IEEE Transactions on Neural Networks and Learning Systems*, 32(6):2373–2387, 2020.

Hong-Ming Chiu, Hao Chen, Huan Zhang, and Richard Y Zhang. Sdp-crown: Efficient bound propagation for neural network verification with tightness of semidefinite programming. In *Forty-second International Conference on Machine Learning*, 2025.

Adnan Darwiche and Auguste Hirth. On the Reasons Behind Decisions. In *Proc. 24th European Conf. on Artifical Intelligence (ECAI)*, pp. 712–720, 2020.

James Enouen and Yan Liu. InstaSHAP: Interpretable Additive Models Explain Shapley Values Instantly. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

Trevor J Hastie. Generalized additive models. In *Statistical models in S*, pp. 249–307. Routledge, 2017.

Xuanxiang Huang, Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva. On Efficiently Explaining Graph-Based Classifiers. In *Proc. 18th Int. Conf. on Principles of Knowledge Representation and Reasoning (KR)*, 2021.

Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. Abduction-Based Explanations for Machine Learning Models. In *Proc. AAAI Conf. on Artificial Intelligence*, pp. 1511–1519, 2019.

Alexey Ignatiev, Yacine Izza, Peter Stuckey, and Joao Marques-Silva. Using MaxSAT for efficient explanations of tree ensembles. In *Proc. 36th AAAI Conf. on Artificial Intelligence*, pp. 3776–3785, 2022.

Yacine Izza and Joao Marques-Silva. On Explaining Random Forests with SAT. In *Proc. 30th Int. Joint Conf. on Artifical Intelligence (IJCAI)*, 2021.

Yacine Izza, Xuanxiang Huang, Antonio Morgado, Jordi Planes, Alexey Ignatiev, and Joao Marques-Silva. Distance-Restricted Explanations: Theoretical Underpinnings & Efficient implementation. In *Proc. 21st Int. Conf. on Principles of Knowledge Representation and Reasoning (KR)*, pp. 475–486, 2024.

Yining Jiao, Carlton J Zdanski, Julia S Kimbell, Andrew Prince, Cameron Worden, Samuel Kirse, Christopher Rutter, Benjamin Shields, William Dunn, Jisan Mahmud, et al. NAISR: A 3D Neural Additive Model for Interpretable Shape Representation. In *International Conference on Learning Representations (ICLR)*, 2024.

Guy Katz, Clark Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer. Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks. In *Proc. 29th Int. Conf. on Computer Aided Verification (CAV)*, pp. 97–117, 2017.

Young Kyung Kim, Juan Matias Di Martino, and Guillermo Sapiro. Generalizing Neural Additive Models via Statistical Multimodal Analysis. *Transactions on Machine Learning Research*, 2024.

Suhas Kotha, Christopher Brix, J. Zico Kolter, Krishnamurthy Dvijotham, and Huan Zhang. Provably bounding neural network preimages. volume 36, pp. 80270–80290, 2023.

Emanuele La Malfa, Agnieszka Zbrzezny, Rhiannon Michelmore, Nicola Paoletti, and Marta Kwiatkowska. On Guaranteed Optimal Robust Explanations for NLP Models. In *Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI)*, pp. 2658–2665, 2021.

Jiachang Liu, Chudi Zhong, Margo Seltzer, and Cynthia Rudin. Fast Sparse Classification for Generalized Linear and Additive Models. *Proceedings of machine learning research*, 151:9304, 2022.

Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Proc. 30th Int. Conf. on Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

J. Marques-Silva, T. Gerspacher, M. Cooper, A. Ignatiev, and N. Narodytska. Explaining Naive Bayes and Other Linear Classifiers with Polynomial Time and Delay. In *Proc. 33rd Int. Conf. on Advances in Neural Information Processing Systems (NeurIPS)*, pp. 20590–20600, 2020.

Joao Marques-Silva. Logic-Based Explainability in Machine Learning. In *Reasoning Web. Causality, Explanations and Declarative Knowledge: 18th Int. Summer School 2022, Berlin, Germany, September 27–30, 2022, Tutorial Lectures*, pp. 24–104. 2023.

Joao Marques-Silva and Alexey Ignatiev. Delivering Trustworthy AI Through Formal XAI. In *Proc. 36th AAAI Conf. on Artificial Intelligence*, pp. 12342–12350, 2022.

Joao Marques-Silva, Thomas Gerspacher, Martin Cooper, Alexey Ignatiev, and Nina Narodytska. Explanations for monotonic classifiers. In *Proc. 38th Int. Conf. on Machine Learning (ICML)*, pp. 7469–7479, 2021.

John Ashworth Nelder and Robert WM Wedderburn. Generalized Linear Models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3):370–384, 1972.

Filip Radenovic, Abhimanyu Dubey, and Dhruv Mahajan. Neural Basis Models for Interpretability. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:8414–8426, 2022.

M. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?" Explaining the predictions of any classifier. In *Proc. 22nd Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 1135–1144, 2016.

M. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-Precision Model-Agnostic Explanations. In *Proc. 32nd AAAI Conf. on Artificial Intelligence*, 2018.

Marco Sälzer and Martin Lange. Reachability is NP-Complete even for the Simplest Neural Networks. In *Proc. 15th Int. Conf. on Reachability Problems (RP)*, pp. 149–164, 2021.

Bernardo Subercaseaux, Marcelo Arenas, and Kuldeep Meel. Probabilistic Explanations for Linear Models. In *Proc. 39th AAAI Conference on Artificial Intelligence*, pp. 20655–20662, 2025.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *Proc. 34th Int. Conf. on Machine Learning (ICML)*, pp. 3319–3328, 2017.

Anton Frederik Thielmann, René-Marcel Kruse, Thomas Kneib, and Benjamin Säfken. Neural additive models for location scale and shape: A framework for interpretable neural regression beyond the mean. In *International Conference on Artificial Intelligence and Statistics*, pp. 1783–1791, 2024.

Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J. Zico Kolter. Beta-Crown: Efficient Bound Propagation with Per-Neuron Split Constraints for Neural Network Robustness Verification. In *Proc. 34th Int. Conf. on Advances in Neural Information Processing Systems (NeurIPS)*, pp. 29909–29921, 2021.

Min Wu, Haoze Wu, and Clark Barrett. Verix: Towards Verified Explainability of Deep Neural Networks. *Proc. 36th Int. Conf. on Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Min Wu, Xiaofu Li, Haoze Wu, and Clark Barrett. Better Verified Explanations with Applications to Incorrectness and Out-of-Distribution Detection. *arXiv preprint arXiv:2409.03060*, 2024.

Shiyun Xu, Zhiqi Bu, Pratik Chaudhari, and Ian J Barnett. Sparse neural additive model: Interpretable deep learning with feature selection via group sparsity. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 343–359, 2023.

Wei Zhang, Brian Barr, and John Paisley. Gaussian Process Neural Additive Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16865–16872, 2024.

Chudi Zhong, Zhi Chen, Jiachang Liu, Margo Seltzer, and Cynthia Rudin. Exploring and Interacting with the Set of Good Sparse Generalized Additive Models. *Advances in neural information processing systems (NeurIPS)*, 36:56673–56699, 2023.

Duo Zhou, Christopher Brix, Grani A Hanasusanto, and Huan Zhang. Scalable neural network verification with branch-and-bound inferred cutting planes. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Appendix

The appendix contains all proofs, optimizations, additional settings, and additional experiments that were mentioned throughout the paper:

Appendix A contains the proofs of Prop. 1 to 4.
Appendix B contains a theoretical discussion and practical optimizations on the importance sorting.
Appendix C contains extensions to multi-class classification and regression tasks.
Appendix D contains all experimental details and ablation studies.
Appendix E contains an LLM usage disclosure.

A Proofs

A.1 Proof of Proposition 1

Proposition 1. *Given a NAM f, an input $\boldsymbol{x} \in \mathbb{R}^n$ and a perturbation radius $\epsilon_p \in \mathbb{R}_+$, let Alg. 2 return a total list order over the input features according to their importance. Then, the following holds: for any sufficient explanation $\mathcal{S}$ that includes feature i, and for any feature $j \notin \mathcal{S}$ such that $i \prec j$ in the list ordering, the set $\mathcal{S} \setminus \{i\} \cup \{j\}$ is also a sufficient explanation.*

Proof. We recall that we have assumed in Alg. 2 that $f(\mathbf{x})$ yields a positive prediction, i.e., it is classified as 1. Accordingly, the final list of bounds $[(\hat{\Delta l}_1, \hat{\Delta u}_1), (\hat{\Delta l}_2, \hat{\Delta u}_2), \ldots, (\hat{\Delta l}_n, \hat{\Delta u}_n)]$ is derived by taking the minimum possible value of each $f_i(\tilde{\mathbf{x}}_i)$ and computing lower and upper bounds for $f_i(\mathbf{x}_{(i)}) - f_i(\tilde{\mathbf{x}}_i)$. The proof we present applies symmetrically to the case where the prediction is negative: in that case, we instead take the maximum value of $f_i(\tilde{\mathbf{x}}_i)$ and bound $f_i(\tilde{\mathbf{x}}_i) - f_i(\mathbf{x}_{(i)})$. We defer a detailed discussion of that case to later. Since we are assuming $f(\mathbf{x}) \geq 0$, the condition that $\mathcal{S}$ is a sufficient explanation with respect to $\langle f, \mathbf{x}, \epsilon_p \rangle$ means that:

$$\begin{aligned} \forall \tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x}). \quad f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) \geq 0 \iff \\ \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) \geq 0. \end{aligned} \tag{3}$$

The value of $f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}})$ is obtained by fixing the features in $\mathcal{S}$ to $\mathbf{x}$ and perturbing the complementary features $\bar{\mathcal{S}}$ to values from $\tilde{\mathbf{x}}$. Owing to this construction, and to the additive form of the NAM f, which can be expressed as $f(\mathbf{x}) := \sum_{t \in [n]} f_t(\mathbf{x}_t)$, we can establish the following statements:

$$\begin{aligned} \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta l}_t \leq (f_t(\mathbf{x}_t) - \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} \sum_{t \in \bar{\mathcal{S}}} f_t(\tilde{\mathbf{x}}_t)) \leq \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta u}_t \iff \\ \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta l}_t \leq \sum_{t \in \bar{\mathcal{S}}} (f_t(\mathbf{x}_t) - \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f_t(\tilde{\mathbf{x}}_t)) + \sum_{t \in \mathcal{S}} (f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t)) \leq \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta u}_t \iff \\ \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta l}_t \leq \sum_{t \in [n]} f(\mathbf{x}) - \sum_{t \in \mathcal{S}} f(\mathbf{x}_t) - \sum_{t \in \bar{\mathcal{S}}} \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\tilde{\mathbf{x}}_t) \leq \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta u}_t \iff \\ \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta l}_t \leq (f(\mathbf{x}) - \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}})) \leq \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta u}_t. \end{aligned} \tag{4}$$

Given our earlier assumption in Equation 3, we know that $\sum_{t \in \bar{\mathcal{S}}} \hat{\Delta u}_t \geq 0$. Now define $\mathcal{S}' := \mathcal{S} \cup \{j\} \setminus \{i\}$. By applying the same line of reasoning as before, we obtain:

$$\begin{aligned} \sum_{t \in \bar{\mathcal{S}}'} \hat{\Delta l}_t \leq (f(\mathbf{x}) - \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}'})) \leq \sum_{t \in \bar{\mathcal{S}}'} \hat{\Delta u}_t \iff \\ \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta l}_t + \hat{\Delta l}_j - \hat{\Delta l}_i \leq (f(\mathbf{x}) - \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}'})) \leq \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta u}_t + \hat{\Delta u}_j - \hat{\Delta u}_i. \end{aligned} \tag{5}$$

Since we assume that $i \prec j$ in the ordering of $[(\hat{\Delta}l_1, \hat{\Delta}u_1), (\hat{\Delta}l_2, \hat{\Delta}u_2), \ldots, (\hat{\Delta}l_n, \hat{\Delta}u_n)]$ and that the bounds are *non-intersecting*, it follows that $\hat{\Delta}l_j - \hat{\Delta}l_i \geq 0$ and $\hat{\Delta}u_j - \hat{\Delta}u_i \geq 0$. Consequently, we obtain that $(f(\mathbf{x}) - \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}'}))$ is bounded both above and below by smaller values than $(f(\mathbf{x}) - \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}}))$. This in turn implies that:

$$\begin{gathered}(f(\mathbf{x}) - \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}'})) - (f(\mathbf{x}) - \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}})) \leq 0 \iff \\ \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) - \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}'}) \leq 0.\end{gathered} \tag{6}$$

Moreover, since $\min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}})$ is non-negative by Equation 3, it follows that:

$$\begin{gathered}\min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}'}) \geq \min_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) \geq 0 \implies \\ \forall \tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x}). \quad f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}'}) \geq 0.\end{gathered} \tag{7}$$

which establishes that $\mathcal{S}'$ constitutes a sufficient explanation for $\langle f, \mathbf{x}, \epsilon_p \rangle$, thereby concluding this part of the proof.

We now turn to the symmetric case, where $f(\mathbf{x}) < 0$. In this setting, Alg. 2 is applied symmetrically by taking the *maximum* admissible value of each $f_i(\tilde{\mathbf{x}}_i)$ and deriving corresponding upper and lower bounds for $f_i(\tilde{\mathbf{x}}_i) - f_i(\mathbf{x}_{(i)})$. $[(\hat{\Delta}l_1, \hat{\Delta}u_1), (\hat{\Delta}l_2, \hat{\Delta}u_2), \ldots, (\hat{\Delta}l_n, \hat{\Delta}u_n)]$ is now sorted in descending importance values, instead of ascending. Given the assumption that $f(\mathbf{x}) < 0$, the requirement that $\mathcal{S}$ constitutes a sufficient explanation with respect to $\langle f, \mathbf{x}, \epsilon_p \rangle$ can be expressed as:

$$\begin{gathered}\forall \tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x}). \quad f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) < 0 \iff \\ \max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) < 0.\end{gathered} \tag{8}$$

Analogous to the earlier case, leveraging the additive structure of the NAM f, which can be written as $f(\mathbf{x}) := \sum_{t \in [n]} f_t(\mathbf{x}_t)$, together with the definitions of $f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}})$ and of the bounds $\hat{\Delta}l_i$ and $\hat{\Delta}u_i$, we can derive the following chain of statements:

$$\begin{aligned}\sum_{t \in \bar{\mathcal{S}}} \hat{\Delta}l_t \leq \sum_{t \in \bar{\mathcal{S}}} (\max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f_t(\tilde{\mathbf{x}}_t) - f_t(\mathbf{x}_t)) \leq \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta}u_t \iff \\ \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta}l_t \leq \sum_{t \in \bar{\mathcal{S}}} (\max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f_t(\tilde{\mathbf{x}}_t) - f_t(\mathbf{x}_t)) + \sum_{t \in \mathcal{S}} (f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t)) \leq \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta}u_t \iff \\ \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta}l_t \leq \sum_{t \in \bar{\mathcal{S}}} \max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f_t(\tilde{\mathbf{x}}_t) + \sum_{t \in \mathcal{S}} f_t(\mathbf{x}_t) - \sum_{t \in [n]} f_t(\mathbf{x}_t) \leq \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta}u_t \iff \\ \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta}l_t \leq (\max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) - f(\mathbf{x})) \leq \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta}u_t.\end{aligned} \tag{9}$$

Since we know that $\sum_{t \in \bar{\mathcal{S}}} \hat{\Delta}u_t \geq 0$, and by defining $\mathcal{S}' := \mathcal{S} \cup \{j\} \setminus \{i\}$ as before, we can now derive that:

$$\begin{aligned}\sum_{t \in \bar{\mathcal{S}}'} \hat{\Delta}l_t \leq (\max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}'}) - f(\mathbf{x})) \leq \sum_{t \in \bar{\mathcal{S}}'} \hat{\Delta}u_t \iff \\ \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta}l_t + \hat{\Delta}l_i - \hat{\Delta}l_j \leq (\max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}'}) - f(\mathbf{x})) \leq \sum_{t \in \bar{\mathcal{S}}} \hat{\Delta}u_t + \hat{\Delta}u_i - \hat{\Delta}u_j.\end{aligned} \tag{10}$$

As before, since we assume $i \prec j$ in the ordering of $[(\hat{\Delta}l_1, \hat{\Delta}u_1), (\hat{\Delta}l_2, \hat{\Delta}u_2), \ldots, (\hat{\Delta}l_n, \hat{\Delta}u_n)]$, and given that the bounds are non-intersecting, together with our assumption that this list is sorted by *decreasing* values, it follows that $\hat{\Delta}l_i - \hat{\Delta}l_j \geq 0$ and $\hat{\Delta}u_i - \hat{\Delta}u_j \geq 0$. Consequently, we obtain

a different outcome: namely, $\max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} \big(f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}'}) - f(\mathbf{x})\big)$ is bounded above and below by strictly *larger* values than $\max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} \big(f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) - f(\mathbf{x})\big)$. This in turn implies that:

$$
\begin{aligned}
(\max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}'}) - f(\mathbf{x})) - (\max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) - f(\mathbf{x})) \leq 0 \iff \\
\max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}'}) - \max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) \leq 0.
\end{aligned} \tag{11}
$$

Moreover, since Equation 8 ensures that $\max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}})$ is negative, we obtain:

$$
\begin{aligned}
\max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}'}) \leq \max_{\tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x})} f(\mathbf{x}_{\mathcal{S}}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}}) < 0 \implies \\
\forall \tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x}). \quad f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}'}) < 0.
\end{aligned} \tag{12}
$$

This establishes that $\mathcal{S}'$ is a sufficient explanation for $\langle f, \mathbf{x}, \epsilon_p \rangle$. With this, the negative case for $f(\mathbf{x})$ is resolved, and together with the positive case, the proof is complete.

□

A.2 Proof of Proposition 2

Proposition 2. *Given a NAM f, an input $\boldsymbol{x} \in \mathbb{R}^n$, and a perturbation radius $\epsilon_p \in \mathbb{R}_+$, Alg. 3 performs $\mathcal{O}(n)$ queries and returns a* cardinal-minimal *sufficient explanation. This stands in contrast to Alg. 1, which is only guaranteed to return a* subset minimal *sufficient explanation.*

Proof. We begin by noting that the algorithm proceeds iteratively, making $|n|$ calls to the query $\text{suff}(f, \mathbf{x}, \mathcal{S} \setminus \{i\}, \epsilon_p)$. Each such query can be encoded using a neural network verifier, which implies that the algorithm requires $\mathcal{O}(n)$ invocations in total. We now turn to proving that Alg. 3 indeed produces a cardinal-minimal sufficient explanation with respect to $\langle f, \mathbf{x}, \epsilon_p \rangle$. First, let us prove that Alg. 3 provides a valid sufficient explanation. This result is straightforward since the last condition that is checked is that: $\text{suff}(f, \mathbf{x}, \mathcal{S} \setminus \{i\}, \epsilon_p)$, and after this condition is met $\mathcal{S}$ is updated to be $\mathcal{S} \setminus \{i\}$ and is returned. Hence, by definition, the sufficiency of the returned subset is satisfied.

We will now demonstrate that the generated set S is a cardinal-minimal sufficient explanation with respect to $\langle f, \mathbf{x}, \epsilon_p \rangle$. Let $1 \leq \ell \leq n$ represent the last feature added to $\mathcal{S}$ in line **??** of Alg. 3. Then, for $\mathcal{S}' := \mathcal{S} \setminus \{\ell\}$, it follows that: $\text{suff}(f, \mathbf{x}, \mathcal{S}', \epsilon_p)$ does not hold true, implying that $\mathcal{S}'$ is *not* a sufficient explanation for $\langle f, \mathbf{x}, \epsilon_p \rangle$. We begin by proving a first lemma that will help us proving our proposition:

Lemma 1. *Given a NAM f, let Alg. 2 return the sorted, non-intersecting list of pairs: $[(\hat{\Delta}l_1, \hat{\Delta}u_1), (\hat{\Delta}l_2, \hat{\Delta}u_2), \ldots, (\hat{\Delta}l_n, \hat{\Delta}u_n)]$. Then, the following holds: if $\mathcal{S}$ that denotes the top $|\mathcal{S}|$ features ordered by $[(\hat{\Delta}l_1, \hat{\Delta}u_1), (\hat{\Delta}l_2, \hat{\Delta}u_2), \ldots, (\hat{\Delta}l_n, \hat{\Delta}u_n)]$ is not a sufficient explanation concerning $\langle f, \boldsymbol{x}, \epsilon_p \rangle$, then any subset $\mathcal{S}' \subseteq [n]$ of size $|\mathcal{S}|$ is also not a sufficient explanation concerning $\langle f, \boldsymbol{x}, \epsilon_p \rangle$.*

Proof. We begin by noting that $\mathcal{S} \subseteq [n]$ is *not* a sufficient explanation with respect to $\langle f, \mathbf{x}, \epsilon_p \rangle$. Assume, for contradiction, that there exists some $\mathcal{S}' \neq \mathcal{S}$ of the same cardinality as $\mathcal{S}$ that *is* a sufficient explanation with respect to $\langle f, \mathbf{x}, \epsilon_p \rangle$. Since both $\mathcal{S}$ and $\mathcal{S}'$ have equal size, we can map each feature in $\mathcal{S}'$ with one in $\mathcal{S}$ according to their position in the ordering $[(\hat{\Delta}l_1, \hat{\Delta}u_1), (\hat{\Delta}l_2, \hat{\Delta}u_2), \ldots, (\hat{\Delta}l_n, \hat{\Delta}u_n)]$.

By definition, $\mathcal{S}$ consists of the top $|\mathcal{S}|$ features in this ordering. Consequently, under the mapping, each feature in $\mathcal{S}'$ is mapped to a feature of strictly higher or equal rank in $\mathcal{S}$. Now consider a sequence of replacements: at each step, replace a feature of $\mathcal{S}'$ with its corresponding equivalent or higher-ranked feature from $\mathcal{S}$. Prop. 1 ensures that each such replacement preserves sufficiency, since a "lower-ranked" feature is being swapped for a "higher-ranked" one. Iterating this process eventually transforms $\mathcal{S}'$ into $\mathcal{S}$, while preserving sufficiency throughout. Thus, $\mathcal{S}$ must itself be a sufficient explanation with respect to $\langle f, \mathbf{x}, \epsilon_p \rangle$ – contradicting the initial assumption that it is not. This completes the proof.

□

From Lemma 1, since the features in $\mathcal{S}' \setminus \{\ell\}$ are the features with the highest $|\mathcal{S}'| = |\mathcal{S}| - 1$ orderings, it holds that any subset $\mathcal{S}'' \subseteq [n]$ of size $|\mathcal{S}'|$ is not a sufficient explanation. To conclude the remaining parts of our proof, we now will make use of another lemma:

Lemma 2. *Let there be some f, $\boldsymbol{x}$, and ϵ_p. Then if $S \in [n]$ is not a sufficient explanation concerning $\langle f, \boldsymbol{x}, \epsilon_p \rangle$, then any $S' \subseteq S$ is not a sufficient explanation w.r.t $\langle f, \boldsymbol{x}, \epsilon_p \rangle$.*

Proof. If $\mathcal{S} \subseteq [n]$ is not a sufficient explanation with respect to $\langle f, \mathbf{x}, \epsilon_p \rangle$, then:

$$\exists \mathbf{z} \in B_p^{\epsilon_p}(\mathbf{x}). \quad f(\mathbf{x}_{\mathcal{S}}; \mathbf{z}_{\bar{\mathcal{S}}}) \neq f(\mathbf{x}). \tag{13}$$

Assume, towards contradiction, that there exists some $\mathcal{S}' \subseteq \mathcal{S}$ which is a sufficient explanation. In other words:

$$\forall \tilde{\mathbf{x}} \in B_p^{\epsilon_p}(\mathbf{x}). \quad f(\mathbf{x}_{\mathcal{S}'}; \tilde{\mathbf{x}}_{\bar{\mathcal{S}}'}) = f(\mathbf{x}). \tag{14}$$

However, consider a vector $\mathbf{z}'$ obtained by fixing the features in $\mathcal{S}'$ to $\mathbf{x}$, the features in $\mathcal{S} \setminus \mathcal{S}'$ also to $\mathbf{x}$, and setting all remaining coordinates according to $\mathbf{z}$. By the earlier implication from Equation 14, we must then have that: $f(\mathbf{x}_{\mathcal{S}'}; \mathbf{z}'_{\bar{\mathcal{S}}'}) \neq f(\mathbf{x})$ and this contradicts the assumption that $\mathcal{S}'$ is sufficient (Equation 13).

□

We now proceed with the remaining part of proving our proposition. Since we know that there is no explanation of size $|\mathcal{S}'| = |\mathcal{S}| - 1$ concerning $\langle f, \mathbf{x}, \epsilon_p \rangle$ from the previous part of the proof, we now can use the result in Lemma 2 to conclude that none of the subsets of these subsets of size $|\mathcal{S}'|$ is not a sufficient explanation too, which implies that there does not exist any explanation of size *lower or equal* to $|\mathcal{S}'| - 1$ which is a sufficient explanation of $\langle f, \mathbf{x}, \epsilon_p \rangle$, which proves that $\mathcal{S}$ is a cardinal-minimal sufficient explanation, hence concluding the proof.

□

A.3 Proof of Proposition 3

Proposition 3. *Given ρ parallelized processors, Alg. 2, performs an overall number of $T_p(n) = \mathcal{O}\Big(\big(\frac{n}{p}\big)\log\big(\max_{i\in[n]}(\frac{\beta_i-\alpha_i}{\xi_i})\big)\Big) \xrightarrow[p\to n]{} \mathcal{O}\Big(\log\big(\max_{i\in[n]}(\frac{\beta_i-\alpha_i}{\xi_i})\big)\Big)$ calls to the verifier, each on a $f_i(\cdot)$ component, where $\xi_i := \min\{|\hat{\Delta}l_{i+1} - \hat{\Delta}u_i|, |\hat{\Delta}l_i - \hat{\Delta}u_{i-1}|\}$.*

Proof. The algorithm terminates once no two univariate functions f_i and f_j have overlapping bounds. Because the procedure relies on binary search, each phase divides the current interval into two. Initially, the gap between the upper and lower bounds for a feature i is exactly $\alpha_i - f_i(\mathbf{x}_{(i)})$. The precision achieved for feature i is limited by the smaller of the two distances: either the distance to the bound of the feature directly above it in the ordering $(i+1)$ or the one directly below it $(i-1)$. Accordingly, we denote the overall precision for feature i by ξ_i as:

$$\xi_i := \min\{|\hat{\Delta}l_{i+1} - \hat{\Delta}u_i|, |\hat{\Delta}l_i - \hat{\Delta}u_{i-1}|\}. \tag{15}$$

Overall, given the binary-search procedure, where the interval is split at each iteration, we define the number of splits k_i performed for a single feature i as:

$$\begin{aligned} \frac{\beta_i - \alpha_i}{2^{k_i}} \leq \xi_i \iff \\ k_i \leq \mathcal{O}\big(\log(\frac{\beta_i - \alpha_i}{\xi_i})\big). \end{aligned} \tag{16}$$

Consequently, the feature on which the maximum number of splits is carried out, denoted by $k_{\max}$, is:

$$k_{max} \leq \mathcal{O}\Big(\max_{i\in[n]}\big(\log(\frac{\beta_i-\alpha_i}{\xi_i})\big)\Big). \tag{17}$$

Each feature $i \in [n]$ is therefore bounded by at most k_{max} verification queries. Consequently, the total workload is upper bounded by $n \cdot k_{max}$, and when distributed across ρ threads, this yields the following parallelized complexity result $T_p(n)$:

$$T_\rho(n) \leq \mathcal{O}\big((\frac{n}{\rho}) \cdot k_{max}\big) \leq \mathcal{O}\Big(\big(\frac{n}{\rho}\big)\log\big(\max_{i\in[n]}(\frac{\beta_i-\alpha_i}{\xi_i})\big)\Big). \tag{18}$$

This completes the proof.

□

A.4 Proof of Proposition 4

Proposition 4. *Given a NAM f, an input $\boldsymbol{x} \in \mathbb{R}^n$, and a perturbation radius $\epsilon_p \in \mathbb{R}_+$, Alg. 3 performs $\mathcal{O}(\log(n))$ queries and returns a* cardinal-minimal *sufficient explanation.*

Proof. We will show that, given the ordering of features $[(\hat{\Delta}l_1, \hat{\Delta}u_1), (\hat{\Delta}l_2, \hat{\Delta}u_2), \ldots, (\hat{\Delta}l_n, \hat{\Delta}u_n)]$ there exists exactly one index $i \in [n]$ such that $\mathcal{S} = [i+1]$ is a sufficient explanation while $\mathcal{S}' = [i]$ is not. Moreover, this statement holds for any ordering. Our proof follows as a consequence of a lemma closely related to Lemma 2, which we restate and establish below:

Lemma 3. *Let there be some f, $\boldsymbol{x}$, and ϵ_p. Then if $S \in [n]$ is a sufficient explanation concerning $\langle f, \boldsymbol{x}, \epsilon_p\rangle$, then any S' for which $\mathcal{S} \subseteq \mathcal{S}'$ is also a sufficient explanation w.r.t $\langle f, \boldsymbol{x}, \epsilon_p\rangle$.*

Proof. This follows directly from Lemma 2. Since $\mathcal{S}$ is known to be a sufficient explanation, assume for contradiction that there exists some $\mathcal{S}' \subseteq [n]$ with $\mathcal{S} \subseteq \mathcal{S}'$ such that $\mathcal{S}'$ is not a sufficient explanation. By Lemma 2, this would imply that no subset $\mathcal{S}'' \subseteq \mathcal{S}'$ could be a sufficient explanation – contradicting the fact that $\mathcal{S} \subseteq \mathcal{S}'$ is sufficient.

□

To conclude the proof of the proposition, consider iterating over the features sequentially according to their ordering. We begin with the empty set $\emptyset$ and test whether it is a sufficient explanation with respect to $\langle f, \mathbf{x}, \epsilon_p\rangle$. If it is not, we proceed by adding features one at a time: first $\{1\}$, then $\{1, 2\}$, then $\{1, 2, 3\}$, and so forth. Eventually, we encounter some feature $i \in [n]$ such that $[i]$ is sufficient with respect to $\langle f, \mathbf{x}, \epsilon_p\rangle$. By Lemma 3, any $\mathcal{S}$ satisfying $[i] \subseteq \mathcal{S}$ is also sufficient. Hence, the unique transition from insufficiency to sufficiency occurs between $[i-1]$ and $[i]$, and there can be no later index $j > i$ for which $[j]$ reverts to being non-sufficient before becoming sufficient again.

Since we have already established this claim, it follows that the binary search in Alg. 4, which halts upon identifying the first feature i where $[i]$ is sufficient but $[i-1]$ is not, will return the same subset as the iterative "naive" Alg. 3, which incrementally traverses features in a greedy manner and outputs $[i]$. Moreover, Prop. 2 shows that Alg. 3 always converges to a cardinal-minimal explanation. Consequently, Alg. 4 must also converge to this same cardinal-minimal explanation, but with only $\mathcal{O}(\log(n))$ sufficiency checks rather than $\mathcal{O}(n)$. This completes the proof.

□

B On Theory and Practice of Importance Sorting

B.1 Theoretical Limits of Importance Sorting

We observe that, just as in any neural network verification task, where the derived bounds may coincide exactly with the certification constraints, the same phenomenon can arise in our approach. Hence, in principle, ξ_i can be arbitrarily small and, in the corner case of the two networks having the same deviation for the ball centered in a particular $\mathbf{x}_{(i)}$, even zero. In such a case, the bounds intervals never become disjoint using Alg. 2, and the preorder is not resolved into a total order.

However, if the networks do not behave identically, this only happens for finitely many isolated points, hence with zero probability if we consider the input points $\mathbf{x}_{(i)}$ drawn randomly from $\mathbb{R}$. For theoretical purposes, one could thus derive our logarithmic complexity bound for the expected runtime independent of ξ_i, which we refrain from both for presentation reasons and for the practical irrelevance of the corner cases.

For all practical purposes, the required precision, the ξ_i, and the respective timeout for this procedure can be set according to the numerical precision of the verifiers or the machine precision due to floating point arithmetics. Optimizations that partially overcome this limitation are also discussed in Appendix B.2.

B.2 Practical Optimizations for Importance Sorting

While analyzing the behavior of Alg. 2 in practice, we noticed that the algorithm makes unnecessarily many verifier calls in certain edge cases. We briefly mention these here, along with our optimizations. As in Sec. 4.1, we consider the case where $f(x) = 1$, which requires us to find bounds $[l_i, u_i]$ for the minimum value $f_i(x^*_{(i)})$ for each feature $i \in [n]$. The case $f(x) = 1$ follows symmetrically.

Let us first consider the edge case where $l_i = f_i(x^*_{(i)})$. In this case, Alg. 2 cuts the bounds in half in each iteration towards l_i but never reaches it. We commonly saw this behavior in the experiments for irrelevant features where $l_i = f_i(x^*_{(i)}) = 0$ but $u_i > 0$. If multiple features have such bounds $[0, u_i]$, Alg. 2 struggles to find a total order until the timeout is reached, which is unfortunate given it is due to features that barely contribute to the classification. However, we found that a simple trick to overcome this issue is drawing samples $\mathbf{x}'_{(i)}$, either drawn randomly from the domain or use the counterexample returned by the verifier, and potentially reducing the domain by setting $u_i = \min\{u_i, f_i(\mathbf{x}'_{(i)})\}$. This works well in practice for the case $[0, u_i]$ on NAMs with ReLU activations, as $\mathbf{x}'_{(i)}$ just has to hit the same piece-wise linear region containing $\mathbf{x}^*_{(i)}$, which gets mapped to 0 through f_i.

Similarly, if $u_i \approx f(\mathbf{x}_{(i)})^*$, the bounds get cut in half from the other direction. However, in this case, randomly drawing samples no longer resolves it if $u_i \neq 0$, or if non-ReLU activations are applied. Here, it helps to directly test if $u_i - \delta$ can still be reached for some small $\delta \in \mathbb{R}_+$ instead of waiting until l_i converges there. A good heuristic to switch to this test is if the verifier concludes that $f(\mathbf{x}^*_{(i)}) \in [m_i, u_i]$ but is unable to return a counterexample demonstrating this (as the verifier has to hit u_i more or less exactly given it's close proximity to $f(\mathbf{x}^*_{(i)})$).

These two optimizations often reduce the number of verifier calls described in Prop. 2 to very few verifier calls. In practice, we have seen that often only 3 verifier calls per feature are required to determine the total ordering.

C Extensions to additional settings

C.1 Extension to Multi-Class Classification

For multi-class classification, let us assume the winner class $t = f(x) = \arg\max_{j \in [c]} f_j(\mathbf{x})$ (Sec. 2.3). Then, we can distinguish between two cases: (i) *winner-vs.-all* explanations, and (ii) all pair-wise *winner-vs.-one* explanations, where each case can be reduced to a binary classification task, where the new binary network in (i) is given by $\hat{f}(\mathbf{x}) = f_t(\mathbf{x}) - \max_{j \neq t} f_j(\mathbf{x})$, and for (ii) by $\hat{f}_j(\mathbf{x}) = f_t(\mathbf{x}) - f_j(\mathbf{x})$, $j \in [n]$, with the class 1 corresponding to the original winner class t. Then, we can apply Alg. 2 and Alg. 4 to generate the respective explanations.

C.2 Extension to regression

Let us assume we want to find the subset $\mathcal{S}$ that is sufficient to determine that the prediction will always be *larger* than some deviation $\delta \in \mathbb{R}_+$ to the original output (the same result for finding the same guarantee for a prediction which is *smaller* will be symmetricly opposite). The first part of

the algorithm will be identical to the case of binary classification with a *positive* outcome (the other use-case will align with a *negative* outcome).

Algorithm 5 Regression: Parallel Interval Importance Sorting

Input: NAM f, input $\mathbf{x} \in \mathbb{R}^n$, perturbation radius $\epsilon_p \in \mathbb{R}_+$

for each feature $i \in [n]$ *in parallel* **do**
  Extract initial bounds α_i, β_i for $f_i(\tilde{\mathbf{x}}_i)$ such that $\tilde{\mathbf{x}}_{(i)} \in \mathcal{B}_p^{\epsilon_p}(\mathbf{x}_{(i)})$
  $l_i \leftarrow \alpha_i$, $u_i \leftarrow \beta_i$
  while True **do**
    $m_i \leftarrow \frac{l_i + u_i}{2}$
    if verify($\forall \tilde{\mathbf{x}}_{(i)} \in \mathcal{B}_p^{\epsilon_p}(\mathbf{x}_{(i)}),\ f_i(\tilde{\mathbf{x}}_i) \geq m_i$) **then**
      $l_i \leftarrow m_i$
    else
      $u_i \leftarrow m_i$
    end if
    $\Delta l_i \leftarrow f_i(\mathbf{x}_{(i)}) - l_i$; $\Delta u_i \leftarrow f_i(\mathbf{x}_{(i)}) - u_i$
    if for all $i \neq j$ it holds that: $\Delta u_i \geq \Delta l_j$ or $\Delta u_j \geq \Delta l_i$ **then**
      Break
    end if
  end while
end for
return $\arg\text{sort}([(\Delta l_1, \Delta u_1), \ldots, (\Delta l_n, \Delta u_n)])$ in ascending order

Now we move on to the second part of the algorithm for the regression case:

Algorithm 6 Regression: Greedy cardinal-minimal Linear Explanation Search

Input: NAM f, input $\mathbf{x} \in \mathbb{R}^n$, perturbation radius $\epsilon_p \in \mathbb{R}_+$, output deviation $\delta \in \mathbb{R}_+$

$\mathcal{S} \leftarrow [n]$
for each feature $i \in [n]$, ordered by Alg. 5 **do** ▷ suff(f, $\mathbf{x}$, $\mathcal{S}$, δ, ϵ_p) holds
  if suff(f, $\mathbf{x}$, $\mathcal{S} \setminus \{i\}$, δ, ϵ_p) **then**
    $\mathcal{S} \leftarrow \mathcal{S} \setminus \{i\}$
  end if
end for
return $\mathcal{S}$ ▷ $\mathcal{S}$ is a *cardinal-minimal* explanation concerning $\langle f, \mathbf{x}, \delta, \epsilon_p \rangle$

which is the same algorithm we used for binary classification (Alg. 3), but where now the evaluation of *suff* evaluates to checking whether fixing the feature subset S ensures that the output remains larger than the original input by less than δ. Given this results, this can be extended to our logarithmic version (Alg. 4) as well:

Algorithm 7 Regression: Greedy Cardinal-Minimal Logarithmic Explanation Search

Input: NAM f, input $\mathbf{x} \in \mathbb{R}^n$, perturbation radius $\epsilon_p \in \mathbb{R}_+$, output deviation $\delta \in \mathbb{R}_+$

$F \leftarrow$ total order of features (Alg. 5)
$l \leftarrow 1$; $u \leftarrow n$
while $l \neq u$ **do**
  $m \leftarrow \lfloor \frac{l+u}{2} \rfloor$
  if suff(f, $\mathbf{x}$, $\{F[1], \ldots F[m]\}$, δ, ϵ_p) **then**
    $l \leftarrow m$
  else
    $u \leftarrow m - 1$
  end if
end while
$\mathcal{S} \leftarrow \{F[1], \ldots F[m]\}$
return $\mathcal{S}$ ▷ $\mathcal{S}$ is a *cardinal-minimal* explanation concerning $\langle f, \mathbf{x}, \delta, \epsilon_p \rangle$

Table 3: Varying perturbation radius ϵ on all datasets.

	Breast Cancer		CREDIT		FICO HELOC	
ϵ	Size (↓)	Time [s] (↓)	Size (↓)	Time [s] (↓)	Size (↓)	Time [s] (↓)
0.01	4.00±4.24	35.60±1.34	—	—	—	—
0.1	4.45±3.98	115.08±73.48	1.79±1.07	76.76±88.21	2.77±1.62	131.31±141.90
0.2	6.29±3.50	143.81±121.27	2.80±2.07	97.16±34.45	3.88±1.58	136.82±77.97
0.5	9.76±2.75	146.01±64.76	3.76±2.62	132.67±36.76	5.59±1.80	317.92±222.07

D Experimental Details and Ablation Studies

D.1 Dataset and Experimental Details

Datasets. We evaluate our approach on four widely used benchmark datasets (Agarwal et al., 2021; Radenovic et al., 2022). covering both classification and regression tasks. The Breast-Cancer dataset contains 569 samples with 30 numeric features per sample. to classify tumors as malignant or benign. The CREDIT dataset includes $1,000$ samples with 20 attributes each, for assessing loan repayment probability. The FICO HELOC dataset comprises $10,459$ samples with 23 financial and demographic features, for predicting creditworthiness. These datasets collectively allow us to evaluate the performance and robustness of our method across different problem types, input dimensions, and domain characteristics. **Models.** We trained 3 binary classification NAMs on the first three datasets, and a regression NAM on the last model. We follow the standard architectures in (Agarwal et al., 2021; Radenovic et al., 2022), and train for each feature a network with hidden layers of size $(64, 64, 32)$. The accuracy of the models for Breast Cancer, CREDIT, and FICO HELOC are 97.37%, 94.92%, and 69.02%, respectively.

Evaluation. All presented results are averaged over 50 samples with a time out of 600s, perturbations are w.r.t to the ℓ_∞-norm on the normalized input and a perturbation radius $\epsilon = 0.5$ is used if not stated otherwise. We filtered trivial samples where, e.g., all features are returned as an explanation. For the Breast Cancer dataset, we use $\epsilon = 0.01$ for an interesting comparison to the local strategies. Our experiments are running on a Ubuntu 24.04 machine with 64GB RAM and 13th Gen Intel(R) Core(TM) i7-1365U. The CREDIT experiments are run on a Ubuntu 24.04 machine with a Intel(R) Xeon(R) Platinum 8380 CPU @ 2.30GHz and two NVIDIA A100-PCIE with 40GB. If not otherwise specified, all experiments were limited to 32 CPU threads. In figures, we show the median along with a shaded region depicting the $25/75\%$ quantiles.

D.2 Ablating the Perturbation Radius

In formal XAI, the generated explanations heavily depend on the chosen perturbation radius ϵ. In Tab. 3, we show how the size and generation time of the explanation change with varying ϵ. Generally, the explanation size and the generation time increase with ϵ, which is expected as previously obtained minimal explanations indeed become insufficient and the verification queries become harder to solve as ϵ is increased. Please note that for $\epsilon = 0.01$, insufficiently many samples for proper averages were found where the explanation is non-trivial on CREDIT and FICO.

D.3 Number of Processed Features over Time

In this experiment, we demonstrate how our approach – after the initial sorting phase – processes the features much quicker to obtain a (globally) cardinal-minimal explanation, as it only requires a logarithmic number of verification queries to do so. This even outperforms approaches that obtain (locally) subset-minimal approaches, including the time needed to sort the features (Fig. 3).

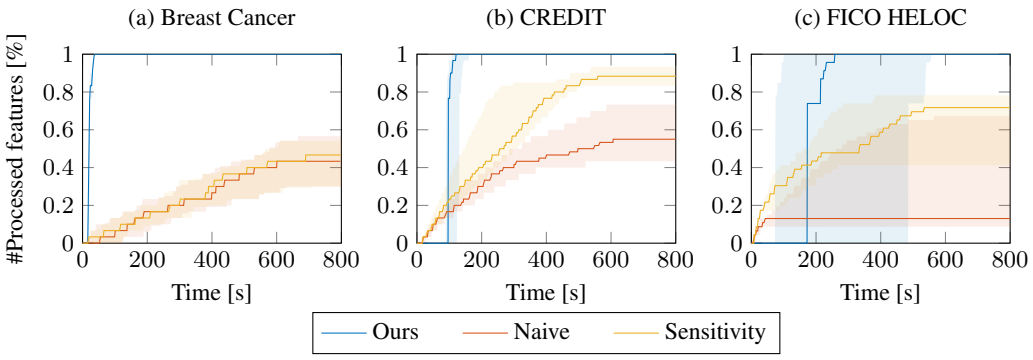

Figure 3: Number of processed features over time.

D.4 Understanding the Parallelization: Ablating the Number of Processors

A key factor influencing the runtime of our approach – particularly the sorting of univariate component importances in Alg. 2 is the *number of processors* allocated for parallelizing the logarithmic binary search. To assess this effect, we conducted an ablation study with varying processor counts.

Fig. 4 illustrates the impact of parallelization on both explanation size and computation time. In particular, the time to sort the features according to their importance (Alg. 2) can be reduced as the number of processors ρ increases, as the bound refinement can be parallelized. In contrast, the subsequent explanation generation (Alg. 4) is barely impacted by the number of processors ρ. Importantly, even with only a single processor (i.e., no parallelization), our algorithm still computes *cardinally*-minimal explanations – a harder task than subset-minimal ones – thus improving explanation size and computation time by design (compare to Fig. 2).

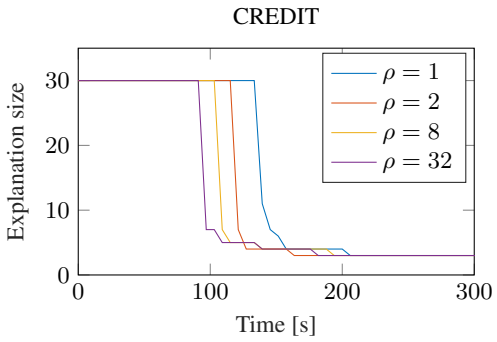

Figure 4: A comparison of explanation generation with different number of processors ρ.

E Disclosure: Use of Large Language Models (LLMs)

A large language model (LLM) was engaged solely as a writing aid to polish language and enhance expression. It played no role in generating research ideas, designing the study, conducting the analysis, or interpreting results. These aspects were performed entirely by the authors.