When predictions are uncertain, can monotonicity help?

Anonymous authors
Paper under double-blind review

Abstract

In recent years, neural networks have achieved impressive performance on a wide range of tasks. However, neural networks tend to make overly optimistic predictions about out-of-distribution data. When managing model risks, it is important to know what we do not know. Although there have been many successes in detecting out-of-distribution data, it is unclear how we can extract further information from these uncertain predictions. To address this problem, we propose to use three types of monotonicity by solving a mean-variance optimization problem. The fast marching method is proposed as an efficient solution. We demonstrate, using empirical examples, that it is possible to provide confident bounds for a large portion of uncertain predictions by monotonicity.

1 Introduction

Deep neural networks (DNNs) have been widely applied in a variety of applications. For high-risk sectors such as the financial sector, predictive power is not the only factor to consider. It is stressed in the model risk management handbook [1] provided by the Office of the Comptroller of the Currency that when machine learning (ML) models are applied, they need to **know what they don't know**. It is possible, for example, for a trading model to perform very well under normal circumstances, but to fail during a financial crisis. To prevent losing money, traders might switch to other strategies if the trading model alerts them to significant changes in the trading environment.

However, in practice, the distribution of the population may differ significantly from that of the training set. For instance, in lending, there is an abundance of data on good applicants, but a very risky applicant with numerous past dues might be extremely rare and have not been seen by the model before. Out-of-distribution (OOD) inputs often pose a challenge to the use of DNNs. Unless carefully managed, DNNs may become overconfident about their predictions, resulting in catastrophic results.

This challenge has received considerable attention in recent years and the results have shown promising (Lakshminarayanan et al., 2017; Kardan et al., 2021; Bibas et al., 2021; Liang et al., 2018; Yang et al., 2022; Geng et al., 2020; Zhou et al., 2022; Ovadia et al., 2019). Based on the successful detection of OOD data, we ask the following question: **What can we learn from what we don't know?**

As with black-box ML models, we may be unable to obtain additional information. On the other hand, recent developments in **domain-knowledge-inspired machine learning models** have been highly successful and could be used to provide additional information. In particular, **monotonic machine learning models** has been very popular in many areas (Repetto, 2022; Chen & Ye, 2022; Liu et al., 2020; Chen & Ye, 2023). There may be additional information to be gained from monotonic models in the case of uncertain predictions. It has been shown by Chen (2022) that individual monotonicity can provide reliable bounds to uncertain inputs. Here is an example of a simple illustration. Let's suppose that an applicant has ten past dues. Although the model may not be certain of this prediction due to its rarity, it does know that five past due has already been very risky, so ten past dues cannot be any less risky. Consequently, it should be categorized into risky groups.

[1] https://www.occ.treas.gov/publications-and-resources/publications/comptrollers-handbook/files/model-risk-management/index-model-risk-management.html

There has been significant attention paid to monotonicity in the past (Sill, 1997; Cano et al., 2019; Gupta et al., 2020), since it is not only about conceptual soundness but also about **fairness**. In the case of credit scoring when individual monotonicity is involved, if an applicant has one more past due, then the model should predict that the probability of default is increasing, otherwise the model would be unfair. Thus, monotonicity is usually a **hard requirement** for related applications.

While individual monotonicity has been extensively studied (Liu et al., 2020; Milani Fard et al., 2016; You et al., 2017; Runje & Shankaranarayana, 2023), **pairwise monotonicity** has been largely ignored. Recent studies (Gupta et al., 2020; Cotter et al., 2019; Chen & Ye, 2022; 2023) have shown that pairwise monotonicity is also important. The idea behind pairwise monotonicity is that some features are intrinsically more important than others. For example, in credit scoring, past due information can be divided into two features based on the number of past dues within three months or more than three months. It is then important to consider the feature that counts the number of past dues of more than three months as more important for fairness. Alternatively, if the ML model predicts an applicant with a past due within three months is more risky than another with a past due more than three months, then the prediction is unfair.

It is possible to provide more information for models containing pairwise monotonicity. For example, if an ML model is sure that an applicant with three past dues between one and three months is already very risky, then an applicant with three past dues greater than three months should only be more risky, even if the model is unsure of the specific predictions it makes. Using this idea, we generalize the mean-variance optimization problem proposed by Chen (2022). This results in a complex **non-convex mixed-integer nonlinear programming** problem. Generally, such a problem is difficult to solve, as discussed in Burer & Letchford (2012). Existing methods may not be capable of solving this problem efficiently. As an example, Chen (2022) does not consider the discrete nature of features and non-convexity. By taking advantage of the monotonic property of models, we propose to use the **fast marching method (FMM)** to find the **global** solution to the problem. The FMM algorithm (Tsitsiklis, 1995; Sethian, 1996; Helmsen et al., 1996) was originally proposed for tracing interface evolution by solving partial differential equations and has been very successful. By utilizing general types of monotonicity, we extend the FMM to solve our optimization problem. By using empirical examples, we demonstrate that our method has the capability of providing reliable bounds to unconfident predictions and enhances the prediction of uncertainty.

In this work, we make three major contributions:

- We generalize the two-stage framework by Chen (2022) with only individual monotonicity to general types of monotonicity. In this way, a larger search space of optimization can be used for the solution, and thus tighter bounds can be provided. Consequently, pairwise monotonicity improves the results.
- The monotonicity-induced optimization geometry of the domain is studied, providing an intuitive understanding of the geometry and permitting the implementation of algorithms in practice.
- A fast marching algorithm based on monotonicity is proposed to find the **global** solution to the complex **non-convex mixed-integer nonlinear programming** optimization.

2 Prerequisites

For problem setup, assume we have n samples and m features, the data-generating process is

$$y_i = f(\mathbf{x}_i) + \epsilon_i \tag{1}$$

for regression problems and

$$y_i|\mathbf{x}_i = \text{Bernoulli}(g^{-1}(f(\mathbf{x}_i))) \tag{2}$$

for binary classification problems, where g is the link function (e.g., logistic function). For simplicity, we assume $x_j \in \mathbb{R}^+ \cup \{0\}$ for all i. Assumptions of this type are common in cost-sharing problems (Friedman & Moulin, 1999) and are often reasonable for high-stakes applications. Then ML methods are applied to approximate f.

2.1 Individual and Pairwise Monotonicity

Monotonicity is crucial for ensuring conceptual soundness and fairness and is therefore often strictly required in high-stakes applications (Chen & Ye, 2023; Gupta et al., 2020; Liu et al., 2020). Throughout the paper, without loss of generality, we focus on the monotonic increasing functions. Suppose f is individual monotonic with respect to x_α and $\neg\boldsymbol{\alpha}$ is the complement of α, then the input $\mathbf{x}$ can be partitioned into $\mathbf{x} = (x_\alpha, \mathbf{x}_{\neg\boldsymbol{\alpha}})$. The well-known **individual monotonicity** is then defined as below.

Definition 2.1. *We say f is **individually monotonic** with respect to x_α if $\forall x_\alpha, x'_\alpha, \mathbf{x}_{\neg\boldsymbol{\alpha}}$,*

$$f(x_\alpha, \mathbf{x}_{\neg\boldsymbol{\alpha}}) \leq f(x'_\alpha, \mathbf{x}_{\neg\boldsymbol{\alpha}}), x_\alpha \leq x'_{\boldsymbol{\alpha}}. \tag{3}$$

In credit scoring, for instance, the probability of default should be individually monotonic with respect to the number of past dues.

In practice, certain features are intrinsically more important than others, and **pairwise monotonicity** (Gupta et al., 2020; Cotter et al., 2019; Chen & Ye, 2023; 2022) describes these phenomena. Analog to equation 3, we partition $\mathbf{x} = (x_\beta, x_\gamma, \mathbf{x}_\neg)$. Without loss of generality, we assume x_β is more important than x_γ. In addition, we require all features with pairwise monotonicity also satisfy individual monotonicity. There are two types of pairwise monotonicity. We start with the **strong pairwise monotonicity**.

Definition 2.2. *We say f is **strongly monotonic** with respect to x_β over x_γ if $\forall x_\beta, x_\gamma, \forall \mathbf{x}_\neg, \forall c \in \mathbb{R}^+$*

$$f(x_\beta, x_\gamma + c, \mathbf{x}_\neg) \leq f(x_\beta + c, x_\gamma, \mathbf{x}_\neg). \tag{4}$$

As an example, in criminology, for each additional crime, a felony is considered more serious than a misdemeanor for predicting the probability of recidivism. Next, we introduce the **weak pairwise monotonicity**.

Definition 2.3. *We say f is **weakly monotonic** with respect to x_β over x_γ if $\forall x_\beta, x_\gamma$ s.t. $x_\beta = x_\gamma, \forall \mathbf{x}_\neg, \forall c \in \mathbb{R}^+$,*

$$f(x_\beta, x_\gamma + c, \mathbf{x}_\neg) \leq f(x_\beta + c, x_\gamma, \mathbf{x}_\neg). \tag{5}$$

When it comes to predicting admission acceptance for STEM majors, math scores on the GRE are more important than verbal scores. In contrast to strong pairwise monotonicity, such comparisons are not always valid. If a student already has a good math score but a very poor verbal score, then an additional increase in verbal scores might increase more chances than math since universities want to ensure the student is capable of communicating effectively. The condition $x_\beta = x_\gamma$ ensures that the comparison is made at the same magnitude since the comparison is not always valid. The result is that, out of 170 points for math and verbal, a student with math 165 and verbal 150 has a greater chance of admission for STEM majors than a student with math 150 and verbal 165, but not necessarily a greater chance than a student with math 160 and verbal 155.

2.2 Detect Out-of-Distribution Data

Lakshminarayanan et al. (2017) describes a simple yet effective approach to detecting OOD data and has shown to be the best performer by Ovadia et al. (2019). Ensemble methods are used. Models are trained M times with random initialization and data shuffles in the entire dataset with $\{\widehat{f}_i(\mathbf{x}; \boldsymbol{\theta}_i)\}$. Models are then used as predictions based on their average,

$$\widehat{\mu}(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^{M} \widehat{f}_i(\mathbf{x}; \boldsymbol{\theta}_i). \tag{6}$$

The variance is used as a proxy for the level of model uncertainty and is calculated by

$$\widehat{\sigma}^2(\mathbf{x}) = \frac{1}{M-1}\sum_{i=1}^{M} (\widehat{f}_i(\mathbf{x}; \boldsymbol{\theta}_i) - \widehat{\mu}(\mathbf{x}))^2. \tag{7}$$

For a data point $\mathbf{x}$, when the variance is large, say $\widehat{\sigma}^2(\mathbf{x}) > \epsilon$ for a threshold ϵ, it is considered too risky to make the prediction. Accordingly, the dataset is divided into confident and unconfident sets.

3 Two-stage Method

We generalize the two-stage framework presented by Chen (2022) for general monotonicity. As a first step, using ensemble methods, we identify the unconfident set for OOD data. In the second stage, we solve a mean-variance optimization problem to provide bounds for the points in the unconfident set. Using general monotonicity, the search domain can be enlarged, resulting in tighter bounds, but also complicating the problem. Lastly, if bounds do not meet expectations, we leave them to human judgment.

3.1 Detect Out-of-Distribution Data

To detect the unconfident set, we wish to utilize the ensemble method. We demonstrate in the following Theorem that monotonicity is preserved by ensembles.

Theorem 3.1. *If $\widehat{f}_i$ achieves all monotonicity for all i, then $\widehat{\mu}$ preserves all monotonicity.*

The proof is left in Appendix A. We then apply the ensemble method and consider

$$\mathbb{S} = \{\mathbf{x} | \widehat{\sigma}^2(\mathbf{x}) \geq \epsilon\} \tag{8}$$

as the **unconfident set**. Similarly, we define $\mathbb{Q} = \{\mathbf{x} | \widehat{\sigma}^2(\mathbf{x}) < \epsilon\}$ as the confident set.

3.2 Mean-Variance Optimization

We wish to provide more information about the unconfident predictions once the unconfident set has been determined. **Without loss of generality, we focus on finding lower bounds, but finding upper bounds is similar.** Suppose we are unconfident about a prediction at $\mathbf{x} \in \mathbb{S}$, we wish to find a confident prediction at $\mathbf{x}'$ with $f(\mathbf{x}') \leq f(\mathbf{x})$ known by monotonicity. Then we have a confident lower bound for $\widehat{\mu}(\mathbf{x})$. Therefore, we wish to maximize $\widehat{\mu}(\mathbf{x}')$ and minimize $\widehat{\sigma}^2(\mathbf{x}')$. However, we cannot perform both optimizations simultaneously, similar to the problems in modern portfolio theory by Markowitz (1952). Hence, we consider only confident prediction with $\widehat{\sigma}^2(\mathbf{x}') < \epsilon$ and look for the largest lower bound $\widehat{\mu}(\mathbf{x}')$. For optimization, we should focus on the domain that can be determined by monotonicity and the following definition is provided.

Definition 3.2. *We define the space $\Omega(\mathbf{x})$ as*

$$\Omega(\mathbf{x}) = \{\mathbf{x}' | f(\mathbf{x}') \underset{M}{\leq} f(\mathbf{x})\}. \tag{9}$$

*whereas $\underset{M}{\leq}$ denotes the **inequality by monotonicity**. That is, we know $f(\mathbf{x}') \leq f(\mathbf{x})$ by the monotonicity from domain knowledge (not from the function outputs).*

In summary, for each $\mathbf{x} \in \mathbb{S}$, we wish to solve the following optimization problem

$$\begin{cases} \max_{\mathbf{x}' \in \Omega(\mathbf{x})} \widehat{\mu}(\mathbf{x}'), \\ \text{subject to } \widehat{\sigma}^2(\mathbf{x}') < \epsilon, \\ \Omega(\mathbf{x}) = \{\mathbf{x}' | f(\mathbf{x}') \underset{M}{\leq} f(\mathbf{x})\}. \end{cases} \tag{10}$$

Although we can obtain the maximum value of $\widehat{\mu}$, it may not be useful if $\widehat{\mu}$ is too small. In practice, we would focus only on $\widehat{\mu} \geq \tau$, for some τ determined by users based on their risk appetites. We may be unable to find satisfactory lower bounds, in which case we leave the decision to human judgment. This may be the case, for example, if we have outliers for nonmonotonic features for which we lack domain expertise. Furthermore, if the variance of all points in the domain is high, there may be no solution. These data points are considered to be part of the **undecided set** $\mathbb{V}$ such that

$$\mathbb{V}(\tau) = \{\mathbf{x} | \mathbf{x} \in \mathbb{S} \text{ and } \min_{\mathbf{x}' \in \Omega(\mathbf{x})} \widehat{\sigma}^2(\mathbf{x}') > \epsilon\} \cup \{\mathbf{x} | \mathbf{x} \in \mathbb{S} \text{ and } \mathbf{x} \text{ solves equation 10 and } \widehat{\mu}(\mathbf{x}) < \tau\}. \tag{11}$$

Similarly, we define the decided set as $\mathbb{W}(\tau) = \{\mathbf{x} | \mathbf{x} \in \mathbb{S} \text{ and } \mathbf{x} \text{ solves equation 10 and } \widehat{\mu}(\mathbf{x}) \geq \tau\}$. A demonstration of the two-stage framework is provided in Figure 1.

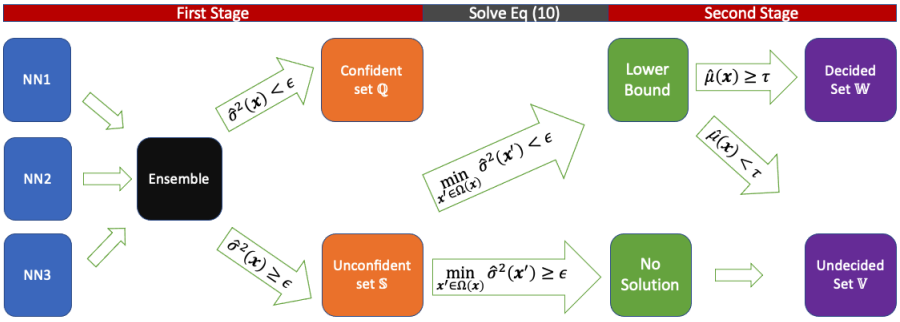


Figure 1: Two-stage framework

4 Geometry of the Domain

We would like to provide a more explicit form for $\Omega(\mathbf{x})$. In the case of only individual monotonicity, $\Omega(\mathbf{x})$ is easily determined as a high-dimensional box. The presence of pairwise monotonicity allows us to have a larger search geometry $\Omega(\mathbf{x})$. However, $\Omega(\mathbf{x})$ is also much more complicated. With this study, we are able to obtain a better understanding of the geometry and also permit us to implement the algorithm in practice.

In the rest of the paper, we will ignore nonmonotonic features for the sake of simplicity, unless otherwise stated, since we cannot draw any conclusions from them. The features are divided into individual monotonic, weak pairwise monotonic, and strong pairwise monotonic parts, as $\mathbf{x} = (\mathbf{x}_S, \mathbf{x}_U, \mathbf{x}_P)$. For features with weak pairwise monotonicity in U, we give two lists $\mathbf{u}$ and $\mathbf{v}$ with $|\mathbf{u}| = |\mathbf{v}|$ such that f is weakly pairwise monotonic with respect to x_{u_i} over x_{v_i} for $i = 1, \ldots, |\mathbf{u}|$. For strong pairwise monotonicity, we assume that there is a list $\mathbf{p}$ such that f is strongly pairwise monotonic to x_{p_j} over $x_{p_{j+1}}$ for $j = 1, \ldots, |\mathbf{p}| - 1$. All monotonic features follow individual monotonicity. This structure is sufficient for most applications, but more complicated structures can be considered if necessary. All proofs are left in Appendix A.

4.1 Individual Monotonicity

In the case that $x_1, \ldots, x_m$ only exhibit individual monotonicity, we would have a box, as shown below.

Proposition 4.1. *Suppose f is individually monotonic with respect to $x_1, \ldots, x_m$, then*

$$\Omega(\mathbf{x}) = \{\mathbf{x}' | \mathbf{x}' \leq \mathbf{x}\} = \{(x_1', \ldots, x_m') | x_1' \leq x_1, \ldots, x_m' \leq x_m\}.$$

Figure 2a with $\mathbf{x} = (3, 1)$ is provided for demonstration. A reduction operator is defined for detecting data points that can be determined by individual monotonicity, which will be used later in the manuscript.

Definition 4.2. *We denote $\psi(\mathbf{x}, \mathbf{c})$ to reduce values of $\mathbf{x}$ by $\mathbf{c}$, that is,*

$$\psi(\mathbf{x}, \mathbf{c}) = \mathbf{x} - \mathbf{c}. \tag{12}$$

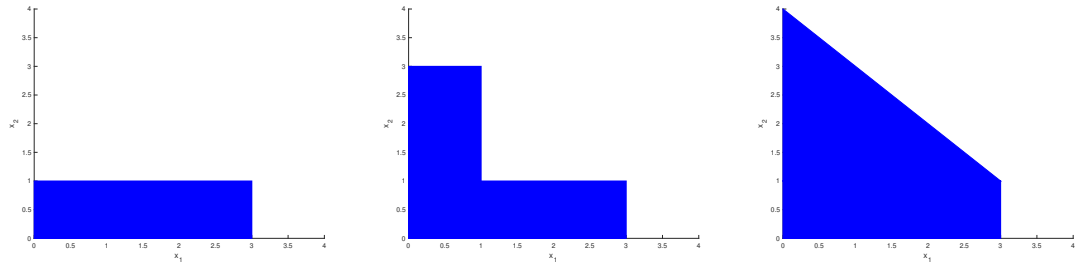
It is easy to check that if $\mathbf{x}' \in \Omega(\mathbf{x})$, then $\exists \mathbf{c}$ such that $\mathbf{x}' = \psi(\mathbf{x}, \mathbf{c})$.

4.2 Weak Pairwise Monotonicity

Pairwise monotonicity presents a more challenging situation. Firstly, we will identify the maximum boundary points, followed by determining the interior points. A swap operator is defined in order to detect data points that can be identified by weak pairwise monotonicity.

Definition 4.3. *For indices β and γ, we denote $\Gamma(\mathbf{x}, \beta, \gamma)$ to swap values of x_β and x_γ in $\mathbf{x}$, that is, if $\mathbf{x}' = \Gamma(\mathbf{x}, \beta, \gamma)$, then*

$$x_i' = \begin{cases} x_i & \text{if } i \neq \beta \text{ and } i \neq \gamma, \\ x_\gamma, & \text{if } i = \beta, \\ x_\beta, & \text{if } i = \gamma. \end{cases} \tag{13}$$



(a) Individual monotonicity (b) Weak pairwise monotonicity (c) Strong Pairwise monotonicity.

Figure 2: Geometry for $\Omega(\mathbf{x})$, where $\mathbf{x} = (x_1, x_2) = (3, 1)$. With the progression from individual monotonicity to weak pairwise monotonicity to strong pairwise monotonicity, we are able to obtain a larger geometry. In the case of weak pairwise monotonicity, the geometry may not be convex.

With the size of $x_\beta + x_\gamma$ fixed, we obtain the following proposition under weak pairwise monotonicity.

Proposition 4.4. *Suppose f is weakly pairwise monotonic with respect to x_β over x_γ, then*

$$f(\Gamma(\mathbf{x}, \beta, \gamma)) \underset{M}{\leq} f(\mathbf{x}), \text{ if } x_\beta > x_\gamma. \tag{14}$$

As a result, we can identify the maximum boundary points of the domain, which we define as follows.

Definition 4.5. *The set of maximum boundary points of Ω is defined as follows:*

$$\partial\Omega(\mathbf{x}) = \left\{ \mathbf{x}' \middle| \sum_{i=1}^{m} x'_i = \sum_{i=1}^{m} x_i, f(\mathbf{x}') \underset{M}{\leq} f(\mathbf{x}) \right\}. \tag{15}$$

We consider swapping by weak pairwise monotonicity to determine maximum boundary points. As a result, we provide the following proposition.

Proposition 4.6. *Suppose f is weakly monotonic with respect to x_{u_i} over x_{v_i} for $i = 1, \ldots, |\mathbf{u}|$, then*

$$\partial\Omega(\mathbf{x}) = \bigcup_{i: x_{u_i} > x_{v_i}} \partial\Omega(\mathbf{x}' | \mathbf{x}' = \Gamma(\mathbf{x}, u_i, v_i)). \tag{16}$$

In other words, by fixing the size of $\sum_{i:i\in U} x_i$, we consider all possible swaps. Recursive definitions are used since swaps can be performed more than once. Clearly, for a new point $\mathbf{x}'$, if $\sum_{i=1}^{m} x'_i > \sum_{i=1}^{m} x_i$, we don't have enough information to compare it with $\mathbf{x}$. Conversely, we have the following theorem.

Theorem 4.7. *Suppose f is weakly monotonic with respect to x_{u_i} over x_{v_i} for $i = 1, \ldots, |\mathbf{u}|$, there exists a $\mathbf{x}'$ with $\sum_{i\in U} x'_i < \sum_{i\in U} x_i$ and $f(\mathbf{x}') \underset{M}{\leq} f(\mathbf{x})$, then there exists a $\widetilde{\mathbf{x}}' \in \partial\Omega(\mathbf{x})$ such that $\mathbf{x}' \leq \widetilde{\mathbf{x}}'$.*

By Theorem 4.7, we need only consider the "interior" of $\partial\Omega$ to determine the domain. It should be noted that in this case, the interior has a different definition.

Definition 4.8. *We say $\mathbf{x}$ is an interior point of $\partial\Omega$ if $\mathbf{x} \leq \mathbf{x}'$ for some $\mathbf{x}' \in \partial\Omega$. Correspondingly, we write $\mathbf{x} \leq \partial\Omega$ if $\mathbf{x} \leq \mathbf{x}'$ for some $\mathbf{x}'$ in $\partial\Omega$.*

We use this definition for better demonstration. If $|u| = |v| = 1$ and $x_u > x_v$, we have

$$\Omega(\mathbf{x}) = \{\mathbf{x}' | \mathbf{x}' \leq \mathbf{x}\} \cup \{\mathbf{x}' | \mathbf{x}' \leq \Gamma(\mathbf{x}, u, v)\}$$

More generally, we have the following proposition, as a result of Proposition 4.6 and Theorem 4.7.

Proposition 4.9. *Suppose f is weakly monotonic with respect to x_{u_i} over x_{v_i}, then*

$$\Omega(\mathbf{x}) = \{\mathbf{x}' | \mathbf{x}' \leq \mathbf{x}\} \cup \bigcup_{i: x_{u_i} > x_{v_i}} \Omega(\mathbf{x}' | \mathbf{x}' = \Gamma(\mathbf{x}, u_i, v_i)). \tag{17}$$

Remark 4.10. *The geometry becomes larger in the presence of weak pairwise monotonicity compared to only individual monotonicity as illustrated in Figure 2b. However, $\Omega(\mathbf{x})$ is not necessarily convex. Suppose, for example, that f is weakly monotonic with respect to x_1 over x_2, and we have $\mathbf{x} = (3, 1)$, then we have*

$$\Omega(\mathbf{x}) = \{(x_1', x_2') | x_1' \leq 3, x_2' \leq 1\} \cup \{(x_1', x_2') | x_1' \leq 1, x_2' \leq 3\}.$$

As shown in Figure 2b, it is evident that it is not convex, and thus complicate the optimization.

4.3 Strong Pairwise Monotonicity

We then consider the strong pairwise monotonicity. Suppose we have a list $\mathbf{p}$ such that f is strongly pairwise monotonic with respect to x_{p_i} over $x_{p_{i+1}}$ for $i = 1, \ldots, |\mathbf{p}| - 1$. As a starting point, we will consider the case in two dimensions. Assume that f is strongly monotonic with respect to x_1 over x_2. Based on the strong pairwise monotonicity of $x_1 + x_2$, we can derive the following proposition.

Proposition 4.11. *Suppose f is strongly pairwise monotonic with respect to x_β over x_γ, if $x_\beta' \leq x_\beta, x_\gamma' = x_\beta + x_\gamma - x_\beta'$, then*

$$f(x_\beta', x_\gamma', \mathbf{x}_\neg) \underset{M}{\leq} f(x_\beta, x_\gamma, \mathbf{x}_\neg). \tag{18}$$

As a result, it can be shown that

$$\Omega(\mathbf{x}) = \{(x_1', x_2') | x_1' \leq x_1, x_2' \leq x_1 + x_2 - x_1'\}.$$

A simple example of $\mathbf{x} = (3, 1)$ is given in Figure 2c with a comparison to individual and weak pairwise monotonicity. Clearly, we would be able to have a larger searching geometry with strong pairwise monotonicity. We consider the following theorem for more general cases.

Theorem 4.12. *For $f(x_1, \ldots, x_m)$, where f is strongly pairwise monotonic with respect to x_i over x_{i+1} for $i = 1, \ldots, m - 1$, then*

$$\Omega(\mathbf{x}) = \left\{ \mathbf{x}' \middle| x_i' \leq \sum_{j=1}^{i} x_j - \sum_{j=1}^{i-1} x_j', \forall i \right\}. \tag{19}$$

As a result of Theorem 4.12, we have the following proposition.

Proposition 4.13. *For $f(x_1, \ldots, x_m)$, where f is strongly pairwise monotonic with respect to x_i over x_{i+1} for $i = 1, \ldots, m - 1$, denote*

$$\varphi(\mathbf{x}, \mathbf{p}) = \left\{ \mathbf{x}' \middle| x_i' \leq \sum_{j=1}^{i} x_j - \sum_{j=1}^{i-1} x_j', \forall i, \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} x_i' \right\}, \tag{20}$$

then

$$\partial\Omega(\mathbf{x}) = \varphi(\mathbf{x}, \mathbf{p}). \tag{21}$$

As a result, if there exists a $\mathbf{x}'$ with $\sum_{i=1}^m x_i' < \sum_{i=1}^m x_i$ and $f(\mathbf{x}') \underset{M}{\leq} f(\mathbf{x})$, then there exists a $\widetilde{\mathbf{x}}' \in \partial\Omega(\mathbf{x})$ such that $\mathbf{x}' \leq \widetilde{\mathbf{x}}'$.

That is, $\mathbf{x}' \in \Omega(\mathbf{x})$ if and only if $\mathbf{x}'$ is interior of $\partial\Omega(\mathbf{x})$.

4.4 General Cases

In general, suppose all features are individual monotonic, f is weakly monotonic with respect to x_{u_i} over x_{v_i} for lists $\mathbf{u}$ and $\mathbf{v}$ in U, and f is strongly monotonic with respect to p_j over p_{j+1} for $j = 1, \ldots, |\mathbf{p}| - 1$. We generalize φ by requiring that $\mathbf{x}_\neg' = \mathbf{x}_\neg$ for features not in $\mathbf{p}$ in equation 20. Then we write the geometry recursively as

$$\Omega(\mathbf{x}) = \{\mathbf{x}' | \mathbf{x}' \leq \mathbf{x}\} \cup \{\mathbf{x}' | \mathbf{x}' \leq \varphi(\mathbf{x}, \mathbf{p})\} \cup \bigcup_{i: x_{u_i} > x_{v_i}} \Omega(\mathbf{x}' | \mathbf{x}' = \Gamma(\mathbf{x}, u_i, v_i)). \tag{22}$$

5 Fast Marching Method

We discuss how to solve the optimization problem equation 10 in this section. The optimization is challenging because of the **nonlinearity** of $\widehat{\mu}(\mathbf{x})$ and $\widehat{\sigma}(\mathbf{x})$, **discrete** and continuous features, and potential **non-convex** geometry (from both $\mathbf{x}' \in \Omega(\mathbf{x})$ and constraints $\widehat{\sigma}^2(\mathbf{x}') < \epsilon$). Therefore, standard optimization algorithms may not be sufficient and difficulties of such problems are discussed by Burer & Letchford (2012). Chen (2022) neglects the discrete nature of some features and potential non-convexity. Based on the monotonicity results studied in Section 4, we would like to pursue a different approach to find a **global** solution.

We begin by binning the features, with discussions in Appendix D. After binning, as a convenience, we assume that $x_j \in \mathbb{Z}^+ \cup \{0\}$, $j = 1, \ldots, m$.

We would like to make use of monotonicity. Specifically, we want to go through the monotonic sequence

$$\begin{cases} \mathbf{x}^1 \to \mathbf{x}^2 \to \ldots, \\ \text{where } \widehat{\mu}(\mathbf{x}^i) \underset{M}{\geq} \widehat{\mu}(\mathbf{x}^{i+1}). \end{cases} \tag{23}$$

and we stop when $\widehat{\sigma}^2(\mathbf{x}^i) < \epsilon$. By brute-force calculation, all possible points in the space must be calculated and sorted, which can be very expensive. Therefore, we intend to carry out this process iteratively.

For each point $\mathbf{x}$, we want to explore its nearby. Define $\mathbf{e}_i$ as

$$(\mathbf{e}_i)_j = \begin{cases} 1, & \text{if } j = i, \\ 0, & \text{otherwise.} \end{cases}$$

Each time we iterate, we explore $\psi(\mathbf{x}, \mathbf{e}_i)$ for all i, that is, we consider decreasing one unit of the feature. Due to Theorem 4.7 and Proposition 4.13, we only need to include maximum boundary points by pairwise monotonicity in the initial set defined recursively as

$$l(\mathbf{x}) = \mathbf{x} \cup \varphi(\mathbf{x}, \mathbf{p}) \cup \bigcup_{i: x_{u_i} > x_{v_i}} l(\mathbf{x}' | \mathbf{x}' = \Gamma(\mathbf{x}, u_i, v_i)). \tag{24}$$

In each iteration, we define our search as follows:

$$\phi(\mathbf{x}) = \bigcup_{i: x_i > 0} \psi(\mathbf{x}, \mathbf{e}_i). \tag{25}$$

As a result, we develop the marching method.

Algorithm 1 (Fast) Marching Method ((F)MM)

1: **Inputs**: $\mathbf{x}$, $\widehat{\mu}(\mathbf{x})$, $\widehat{\sigma}(\mathbf{x})$, and a set l defined in equation 24
2: **Outputs**: Return the **global** solution to equation 10 if exists
3: **while** $\widehat{\sigma}^2(\mathbf{x}) \geq \epsilon$ and $|l| > 0$ **do**
4: $\quad l = l \cup \{\mathbf{x}' | \mathbf{x}' \in \phi(\mathbf{x})$ and $\mathbf{x}'$ has not been visited$\}$
5: $\quad$ Return $\mathbf{x}$ as the element corresponds to maximum $\widehat{\mu}(\mathbf{x}')$ in l and remove it from l (by Heap)
6: **end while**

The most expensive calculation in the marching algorithm is to determine the maximum value in the set l. A straightforward calculation costs $\mathcal{O}(|l|)$. The heap data structure can accelerate such calculations, as discussed by (Tsitsiklis, 1995; Sethian, 1996; Helmsen et al., 1996). This algorithm is known as the **Fast Marching Method (FMM)**, which has been proven to be a highly effective method for tracing interface evolution by solving partial differential equations. As there are more insertions than extract-max operations, we use the Fibonacci heap by Fredman & Tarjan (1987), which has a lower insertion cost than the binary heap. Different from the original FMM, we utilize general monotonicity from domain knowledge.

5.1 Analysis of the Algorithm

The algorithm is now briefly analyzed with the proof left in Appendix A. The following proposition shows that the algorithm marches monotonically.

Proposition 5.1. *MM searches for the solution in a monotonic non-increasing order of $\widehat{\mu}$.*

To ensure that we do not miss any points during the march, we provide the following proposition.

Proposition 5.2. *When MM runs to the end with $l = \{\}$, all points in the domain have been explored.*

Let us suppose that the iteration stops after N steps. Then the space complexity is $\mathcal{O}(mN)$. It is estimated that the amortized time for inserting is $\Theta(1)$ and deleting the maximum key is $\Theta(\log(|l|))$. We further assume that the calculation of $\widehat{\mu}(\mathbf{x})$ and $\widehat{\sigma}(\mathbf{x})$ is C. As a result, at each iteration, the amortized time is $\Theta(Cm + \log(|l|))$. The overall cost of FMM is

$$\text{Cost}_{FMM} = \Theta((Cm + \log(m))N + N\log(N)).$$

As a comparison, the cost of MM is

$$\text{Cost}_{MM} = \mathcal{O}(CmN + mN^2).$$

As a result, the operation of finding the maximum values will be very expensive for large N.

Remark 5.3. *In practice, the most expensive part is CmN, where C depends on the architectures of models.*

6 Empirical Example

We provide examples in finance, criminology, education, life science, and healthcare. The results are summarized in Table 1. In all experiments, the monotonic groves of neural additive models (MGNAMs) proposed by Chen & Ye (2023) are used. For all examples, we use ten models ($M = 10$) for ensembles and threshold $\epsilon = 10^{-3}$. Note models and thresholds are not unique choices and we provide more discussions in Appendix D. We provide detailed analyses for two datasets and leave the remaining examples to the Appendix B.

6.1 Finance - Credit Scoring

We use the Kaggle credit score dataset, Give Me Some Credit (GSMC). Without loss of generality, we let $x_1 - x_3$ denote the number of past dues and their duration: 90+ days, 60-89 days, and 30-59 days. By domain knowledge, the probability of default is strongly monotonic with respect to x_1 over x_2 over x_3. Furthermore, we impose individual monotonicity for monthly income (x_4) and number of dependents (x_5).

6.1.1 Stage I - Detect OOD Data

As the first step, we apply the ensemble method to detect OOD data. Running the experiment with the entire dataset leads to the identification of approximately 2.8% of the data as uncertain samples, which are therefore categorized in the **unconfident set** $\mathbb{S}$. A common example would be an applicant with a high amount of past dues, which is very rare in the dataset. Considering the rarity of this prediction, it makes sense that our model is unconfident about it. The existence of OOD data seems to be commonplace. Consequently, we should exercise caution when applying our models to new situations.

6.1.2 Stage II - Finding Lower Bounds

Next, we use the FMM to solve the mean-variance optimization problem 10. Predictions higher than τ are left in the **undecided set** $\mathbb{V}$. Our analysis takes into account a variety of choices of τ, which can be selected by users according to their risk appetites. It is important to emphasize that while $\tau = 50\%$ is a natural choice for image classification applications, it is not necessarily the best choice for credit scoring applications. Credit scoring aims to accurately predict the probability of default, and 50% is already a very high probability. For accuracy, we calculate $\frac{|\mathbb{V}|}{|\mathbb{S}|} \in [0, 1]$, which is used to determine the ratio of unconfident

Table 1: Summary results for all datasets

DATASETS	AUC (%)	OOD (%)	$\frac{\lvert\mathbb{V}\rvert}{\lvert\mathbb{S}\rvert}$ (%)						MEAN-ITER
			$\tau = 0.5$	0.4	0.3	0.2	0.1	0	
GMSC	85	2.8	77.4	64.5	43.2	15.7	7.9	1.0	29
COMPAS	72	10.4	97.7	86.9	80.1	74.0	72.5	72.2	4
Law	86	10.5	78.9	58.3	39.0	21.1	14.2	14.2	343
Life-Science	68	14.0	75.0	55.0	30.0	10.0	0.0	0.0	9
Mammography	90	52.2	77.8	76.7	76.4	74.1	71.7	70.7	12

samples that cannot be provided with reliable lower bounds. Thus, 0 suggests that confident lower bounds are provided for all unconfident samples, whereas 1 suggests that no confident lower bounds are provided.

Overall results. Our findings are summarized below, based on Table 1. As a result of considering lower τ, we are able to determine more confident predictions, as expected. Based on strong pairwise monotonicity, only 0.22% of the entire dataset is undecided when $\tau = 0.1$. FMM has a mean number of iterations of 29 for each optimization, demonstrating its efficacy. In this regard, our method has proven to be successful.

Comparison using different monotonicity. Further comparisons are made by considering individual, weak pairwise, and strong pairwise monotonicity. It should be noted that in this example, $x_1 - x_3$ exhibits strong pairwise monotonicity, which implies weak pairwise monotonicity. The case of individual monotonicity is similar to the work done by Chen (2022), but the FMM accounts for discrete features and offers global solutions. We focus on samples that are affected by pairwise monotonicity and denote the set that includes these samples as $\mathbb{T}$. It is important to note that not all samples are affected by pairwise monotonicity; for example, $(x_1, x_2, x_3) = (0, 0, 3)$. In such a scenario, the FMM would produce the same result regardless of whether pairwise monotonicity is considered or not, and therefore it is not of interest to us. Similarly, we denote the undecided set out of $\mathbb{T}$ as $\mathbb{U}$. The results are documented in Table 2. Based on the table, it can be seen that pairwise monotonicity provides consistent improvements for different values of τ, suggesting that pairwise monotonicity should be used. The number of iterations is not significantly increased by including pairwise monotonicity, demonstrating the effectiveness of FMM.

A successful example. To better demonstrate how FMM performs, we provide a successful example with

$$\mathbf{x} = \begin{bmatrix} 3 & 1 & 3 & 1 & 2 & 0.96 & 0.38 & 9 & 1 & 40 \end{bmatrix}.$$

The iterations of FMM are recorded in Table 3. By following a monotonic sequence, variance has been reduced to meet the threshold. The reason for this result is that a large number of past dues are rare. The overall number of past dues $(x_1 + x_2 + x_3)$ is greater than 7 in only 232 samples. As the number of past dues decreases, the model becomes more confident in its prediction. Additionally, strong pairwise monotonicity is necessary in order to get from $(x_1, x_2, x_3) = (3, 1, 3)$ to $(x_1, x_2, x_3) = (2, 2, 2)$.

An unsuccessful example. Unfortunately, not all cases can result in positive outcomes. In some cases, it may not be possible to reduce the variance to the desired level. Here is an example,

$$\mathbf{x} = \begin{bmatrix} 0 & 0 & 1 & 1 & 2 & 0.90 & 1.06 & 25 & 10 & 57 \end{bmatrix}.$$

The iterations of FMM are recorded in Table 4. In spite of the fact that FMM has been run to the end and the variance has been significantly reduced, we are still unable to come up with a satisfactory bound. In this example, it appears that x_9 is a large value that is substantially different from the mean value of 1. In fact, there are only 113 samples with $x_9 \geq 10$, which indicates that this feature is quite rare and results in a high level of variance. As x_9 is not considered a monotonic feature, we cannot provide further information. If additional domain knowledge is included, it may be possible to provide further information. For example, local monotonicity might be imposed instead of global monotonicity (the probability of default increases after N loans). This result encourages us to incorporate more domain knowledge into the model.

Table 2: The comparison for the optimization 10 by considering different monotonicity for the GMSC dataset with samples affected by pairwise monotonicity. Pairwise monotonicity improves the result.

INDIVIDUAL			**WEAK**			**STRONG**		
τ	$\frac{\lvert\mathbb{U}\rvert}{\lvert\mathbb{T}\rvert}(\%)$	MEAN-ITER	τ	$\frac{\lvert\mathbb{U}\rvert}{\lvert\mathbb{T}\rvert}(\%)$	MEAN-ITER	τ	$\frac{\lvert\mathbb{U}\rvert}{\lvert\mathbb{T}\rvert}(\%)$	MEAN-ITER
0.5	72.4	15	0.5	71.9	19	0.5	70.0	24
0.4	58.8	16	0.4	57.7	19	0.4	54.6	24
0.3	35.5	16	0.3	33.7	19	0.3	32.3	24
0.2	11.8	18	0.2	9.0	22	0.2	8.8	28
0.1	6.0	19	0.1	4.6	24	0.1	4.6	31
0	3.6	21	0	3.6	27	0	3.6	35

Table 3: A successful example of GMSC.

ITER	x_1	x_2	x_3	x_4	x_5	**MEAN**	**VARIANCE**
0	3	1	3	1	2	0.6071	0.00165
1	3	0	4	1	2	0.6034	0.00159
2	3	1	2	1	2	0.5989	0.00151
3	3	0	3	1	2	0.5935	0.00142
4	3	1	3	1	1	0.5918	0.00174
5	2	2	3	1	2	0.5881	0.00103
6	3	0	4	1	1	0.5880	0.00168
7	3	1	1	1	2	0.5864	0.00135
8	3	1	2	1	1	0.5834	0.00159
9	2	1	4	1	2	0.5828	0.00103
10	2	2	2	1	2	0.5793	0.00092

6.2 Criminology - Recidivism

We present another example with a less satisfactory result and analyze the reason and potential remedy. In criminology, we examine the prediction of recidivism using the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) (Pro, 2016). There are four monotonic features. The overall result is recorded in Table 1 and the comparison of individual and pairwise monotonicity is in Table 5. Although we can obtain reliable bounds for some data, the performance is not as good as that of the GMSC dataset.

Reasons for the less satisfactory result. There is a difficulty encountered when using the FMM to reduce the variance of OOD data. For example, there are only 20% of which, find the confident lower bound when $\tau = 0.3$. Age appears to be a significant factor, as demonstrated in Appendix B, but there is no indication that it is a global monotonic feature. The relationship between age and the prediction variance, however, shows a clear pattern, as in Figure 3. Roughly speaking, model predictions are much less confident for young people. We expect the performance to improve significantly if we can incorporate further domain knowledge regarding age, such as local monotonicity for example. Even so, because we are not experts in the field, we do not impose further limitations on age here to avoid unfair treatment.

7 Limitations and Future Work

In this paper, we demonstrate how to exploit general monotonicity to reduce prediction uncertainty by providing a general optimization framework and a fast marching method. Overall, pairwise monotonicity has enhanced the performance and the fast marching method provides a global solution with reasonable iterations, making it a suitable benchmark for future research. There are some limitations of our current approach, as summarized below. Correspondingly, we propose future research.

Table 4: An unsuccessful example of GMSC.

ITER	x_1	x_2	x_3	x_4	x_5	**MEAN**	**VARIANCE**
0	0	0	1	1	2	0.4182	0.00375
1	0	0	1	1	1	0.4028	0.00365
2	0	0	1	1	0	0.3820	0.00328
3	0	0	1	0	2	0.3770	0.00431
4	0	0	1	0	1	0.3622	0.00414
5	0	0	1	0	0	0.3422	0.00363
6	0	0	0	1	2	0.2335	0.00195
7	0	0	0	1	1	0.2223	0.00181
8	0	0	0	1	0	0.2074	0.00154
9	0	0	0	0	2	0.2043	0.00186
10	0	0	0	0	1	0.1941	0.00171
11	0	0	0	0	0	0.1806	0.00142

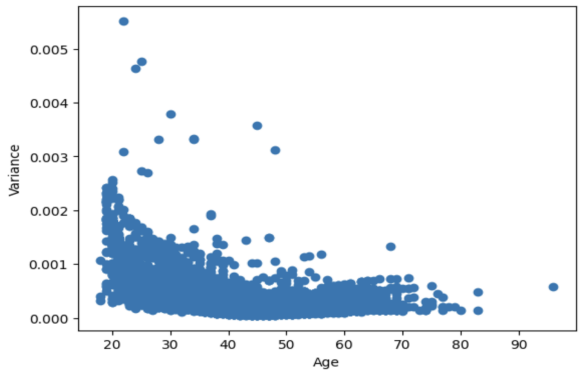


Figure 3: Variance vs Age in the COMPAS dataset.

Table 5: The comparison of the optimization 10 by considering different monotonicity for the COMPAS dataset with samples affected by pairwise monotonicity. Pairwise monotonicity improves the result.

	INDIVIDUAL			**STRONG**	
τ	$\frac{\lvert\mathbb{U}\rvert}{\lvert\mathbb{T}\rvert}$(%)	MEAN-ITER	τ	$\frac{\lvert\mathbb{U}\rvert}{\lvert\mathbb{T}\rvert}$(%)	MEAN-ITER
0.5	73.0	4	0.5	67.6	5
0.4	70.3	7	0.4	64.9	9
0.3	70.3	7	0.3	62.3	10
0.2	67.6	8	0.2	62.3	10
0.1	67.6	8	0.1	62.3	11
0	67.6	8	0	62.3	11

1. The major limitation of the current FMM is that it does not consider the behavior of prediction variance, thus it may take a considerable number of iterations, especially for high-dimensional problems. We plan to utilize properties of model variances to improve the FMM's search process.

2. In general, FMM performs better when more features exhibit monotonicity, especially important features. It should be noted, however, that some important features may not exhibit monotonicity, at least not globally. The performance may be improved by applying other domain knowledge (Gupta et al., 2020; 2018) and imposing local monotonicity. Such a direction will be explored.

References

Propublica. compas data and analysis for "machine bias"., 2016. URL `https://github.com/propublica/compas-analysis.`

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*, 34, 2021.

Koby Bibas, Meir Feder, and Tal Hassner. Single layer predictive normalized maximum likelihood for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:1179–1191, 2021.

Samuel Burer and Adam N Letchford. Non-convex mixed-integer nonlinear programming: A survey. *Surveys in Operations Research and Management Science*, 17(2):97–106, 2012.

Saul Calderon-Ramirez, Diego Murillo-Hernandez, Kevin Rojas-Salazar, Luis-Alexander Calvo-Valverd, Shengxiang Yang, Armaghan Moemeni, David Elizondo, Ezequiel López-Rubio, and Miguel A Molina-Cabello. Improving uncertainty estimations for mammogram classification using semi-supervised learning. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2021.

José-Ramón Cano, Pedro Antonio Gutiérrez, Bartosz Krawczyk, Michał Woźniak, and Salvador García. Monotonic classification: An overview on algorithms, performance measures and data sets. *Neurocomputing*, 341:168–182, 2019.

Dangxing Chen. Two-stage modeling for prediction with confidence. In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 1–5. IEEE, 2022.

Dangxing Chen and Weicheng Ye. Monotonic neural additive models: Pursuing regulated machine learning models for credit scoring. In *Proceedings of the Third ACM International Conference on AI in Finance*, pp. 70–78, 2022.

Dangxing Chen and Weicheng Ye. How to address monotonicity for model risk management? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 5282–5295. PMLR, 23–29 Jul 2023.

Andrew Cotter, Maya Gupta, Heinrich Jiang, Erez Louidor, James Muller, Tamann Narayan, Serena Wang, and Tao Zhu. Shape constraints for set functions. In *International conference on machine learning*, pp. 1388–1396. PMLR, 2019.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580, 2018.

Matthias Elter. Mammographic Mass. UCI Machine Learning Repository, 2007. DOI: https://doi.org/10.24432/C53K6Z.

Michael L Fredman and Robert Endre Tarjan. Fibonacci heaps and their uses in improved network optimization algorithms. *Journal of the ACM (JACM)*, 34(3):596–615, 1987.

Eric Friedman and Herve Moulin. Three methods to share joint costs or surplus. *Journal of economic Theory*, 87(2):275–312, 1999.

Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3614–3631, 2020.

Maya Gupta, Dara Bahri, Andrew Cotter, and Kevin Canini. Diminishing returns shape constraints for interpretability and regularization. *Advances in neural information processing systems*, 31, 2018.

Maya Gupta, Erez Louidor, Oleksandr Mangylov, Nobu Morioka, Taman Narayan, and Sen Zhao. Multidimensional shape constraints. In *International Conference on Machine Learning*, pp. 3918–3928. PMLR, 2020.

John Joseph Helmsen, Elbridge Gerry Puckett, Phillip Colella, and Milo Dorr. Two new methods for simulating photolithography development in 3d. In *Optical Microlithography IX*, volume 2726, pp. 253–261. SPIE, 1996.

Navid Kardan, Ankit Sharma, and Kenneth O Stanley. Towards consistent predictive confidence through fitted ensembles. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9. IEEE, 2021.

Waldemar Koczkodaj. Somerville Happiness Survey. UCI Machine Learning Repository, 2018. DOI: https://doi.org/10.24432/C5PW36.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

Xingchao Liu, Xing Han, Na Zhang, and Qiang Liu. Certified monotonic neural networks. *Advances in Neural Information Processing Systems*, 33:15427–15438, 2020.

Harry Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952. ISSN 00221082, 15406261.

Mahdi Milani Fard, Kevin Canini, Andrew Cotter, Jan Pfeifer, and Maya Gupta. Fast and flexible monotonic functions with ensembles of lattices. *Advances in neural information processing systems*, 29, 2016.

Shayan Shaghayeq Nazari and Pinku Mukherjee. An overview of mammographic density and its association with breast cancer. *Breast cancer*, 25:259–267, 2018.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Arlie O Petters and Xiaoying Dong. An introduction to mathematical finance with applications. *New York, NY: Springer*, 10:978–1, 2016.

Marco Repetto. Multicriteria interpretability driven deep learning. *Annals of Operations Research*, pp. 1–15, 2022.

Davor Runje and Sharath M Shankaranarayana. Constrained monotonic neural networks. In *International Conference on Machine Learning*, pp. 29338–29353. PMLR, 2023.

James A Sethian. A fast marching level set method for monotonically advancing fronts. *proceedings of the National Academy of Sciences*, 93(4):1591–1595, 1996.

Joseph Sill. Monotonic networks. *Advances in neural information processing systems*, 10, 1997.

John N Tsitsiklis. Efficient algorithms for globally optimal trajectories. *IEEE transactions on Automatic Control*, 40(9):1528–1538, 1995.

Linda F Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, WENXUAN PENG, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking generalized out-of-distribution detection. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

Seungil You, David Ding, Kevin Canini, Jan Pfeifer, and Maya Gupta. Deep lattice networks and partial monotonic functions. *Advances in neural information processing systems*, 30, 2017.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

A MISSING PROOFS

This section contains detailed proofs of the results that are missing in the main paper.

A.1 Proof of Theorem 3.1

Proof. Suppose $\widehat{f}_i$ is individually monotonic to x_α, then

$$\widehat{f}_i(x_\alpha, \mathbf{x}_\neg) \underset{M}{\leq} \widehat{f}_i(x'_\alpha, \mathbf{x}_\neg), \text{ if } x_\alpha \leq x'_\alpha.$$

If this is true for all i, then for $x'_\alpha \geq x_\alpha$, we have

$$\frac{1}{M}\sum_{i=1}^{M} \widehat{f}_i(x_\alpha, \mathbf{x}_\neg) \underset{M}{\leq} \frac{1}{M}\sum_{i=1}^{M} \widehat{f}_i(x'_\alpha, \mathbf{x}_\neg)$$

Suppose $\widehat{f}_i$ is weakly pairwise monotonic with respect to x_β over x_γ, then for $x_\beta = x_\gamma$, we have

$$\widehat{f}_i(x_\beta, x_\gamma + c, \mathbf{x}_\neg) \underset{M}{\leq} \widehat{f}_i(x_\beta + c, x_\gamma, \mathbf{x}_\neg), \forall c \in \mathbb{R}^+.$$

If this is true for all i, then we have

$$\frac{1}{M}\sum_{i=1}^{M} \widehat{f}_i(x_\beta, x_\gamma + c, \mathbf{x}_\neg) \underset{M}{\leq} \frac{1}{M}\sum_{i=1}^{M} \widehat{f}_i(x_\beta + c, x_\gamma, \mathbf{x}_\neg).$$

A similar conclusion can be drawn for strong pairwise monotonicity. □

A.2 Proof of Proposition 4.1

Proof. If $\mathbf{x}' \in \Omega(\mathbf{x})$, then $f(\mathbf{x}') \underset{M}{\leq} f(\mathbf{x})$ by definition. Conversely, if $x'_i > x_i$ for some i, we cannot draw any conclusions. □

A.3 Proof of Proposition 4.4

Proof. Without loss of generality, we write $\mathbf{x} = (x_\beta, x_\gamma, \mathbf{x}_\neg)$. Consider $\mathbf{x}' = (x_\gamma, x_\gamma, \mathbf{x}_\neg)$ and $c = x_\beta - x_\gamma > 0$, then by definition, we have

$$f(x_\gamma, x_\gamma + c, \mathbf{x}_\neg) \underset{M}{\leq} f(x_\gamma + c, x_\gamma, \mathbf{x}_\neg).$$

□

A.4 Proof of Proposition 4.6

Proof. From Proposition 4.4, we determine the form. Otherwise, if $x_\beta < x_\gamma$, we cannot draw conclusions. □

A.5 Proof of Theorem 4.7

Proof. Consider the sequence of steps required to make $\mathbf{x}'$ from $\mathbf{x}$,

$$\mathbf{x}^1 \to \mathbf{x}^2 \to \cdots \to \mathbf{x}',$$

with $\mathbf{x}^1 = \mathbf{x}$. For each step, we can either reduce the value of x_i as long as $x'_i \geq 0$, or swap x_i with x_j if $x_i > x_j$. We would like to construct a new sequence by applying reduction operations and swap operations

one by one. In order to accomplish this, we merge all reduction operations between two swap operations. In the absence of a reduction operation, we simply consider $\psi(\mathbf{x}, \mathbf{0})$. As a result, we have

$$\mathbf{x}^{i+1} = \begin{cases} \psi(\mathbf{x}^i, \mathbf{c}^i), & \text{if } i \text{ odd}, \\ \Gamma(\mathbf{x}^i, u^i, v^i), & \text{if } i \text{ even}. \end{cases} \tag{26}$$

Next, we construct another sequence in $\partial\Omega$ to bound $\mathbf{x}^i$,

$$\widetilde{\mathbf{x}}^1 \to \widetilde{\mathbf{x}}^2 \to \cdots \to \widetilde{\mathbf{x}}',$$

with $\widetilde{\mathbf{x}}^1 = \mathbf{x}$ and

$$\widetilde{\mathbf{x}}^{i+1} = \begin{cases} \widetilde{\mathbf{x}}^i, & \text{if } i \text{ odd}, \\ \Gamma(\widetilde{\mathbf{x}}^i, u^i, v^i) & \text{if } i \text{ even and } \widetilde{\mathbf{x}}^i_{u^i} > \widetilde{\mathbf{x}}^i_{v^i}, \\ \widetilde{\mathbf{x}}^i, & \text{if } i \text{ even and } \widetilde{\mathbf{x}}^i_{u^i} < \widetilde{\mathbf{x}}^i_{v^i}. \end{cases}$$

We want to show that $\widetilde{\mathbf{x}}^i \geq \mathbf{x}^i$ for all i. It is clear that this holds for $i = 1$, and we consider when $i > 1$. We focus on the third case because the first two cases are obvious. If $\widetilde{\mathbf{x}}^i_{u^i} < \widetilde{\mathbf{x}}^i_{v^i}$ and $\mathbf{x}^i_{u^i} > \mathbf{x}^i_{v^i}$, since $\mathbf{x}^i_{u^i} \leq \mathbf{x}^{i-1}_{u^{i-1}}$ and $\mathbf{x}^{i-1}_{u^{i-1}} \leq \widetilde{\mathbf{x}}^i_{u^i}$, then $\mathbf{x}^i_{u^i} < \widetilde{\mathbf{x}}^i_{v^i}$ and $\mathbf{x}^i_{v^i} < \widetilde{\mathbf{x}}^i_{u^i}$. Thus, after swapping on $\mathbf{x}^i$, $\mathbf{x}^{i+1} \leq \widetilde{\mathbf{x}}^{i+1}$. By induction, we conclude.

□

A.6 Proof of Proposition 4.11

Proof. Let $c = x_\beta - x'_\beta$, then we have

$$f(x'_\beta, x'_\gamma, \mathbf{x}_\neg) = f(x'_\beta, x_\gamma + c, \mathbf{x}_\neg) \underset{M}{\leq} f(x'_\beta + c, x_\gamma, \mathbf{x}_\neg) = f(x_\beta, x_\gamma, \mathbf{x}_\neg).$$

□

A.7 Proof of Theorem 4.12

Proof. First, we show if $\mathbf{x}' \in \Omega(\mathbf{x})$, then $f(\mathbf{x}') \underset{M}{\leq} f(\mathbf{x})$. Denote $c_i = x_i - x'_i$, then from Equation equation 19 we have

$$\sum_{j=1}^{i} c_j \geq 0, i = 1, \ldots, m.$$

By Proposition 4.11 and individual monotonicity, we have

$$\begin{aligned} f(x_1, \ldots, x_m) &\underset{M}{\geq} f(x'_1, x_2 + c_1, \ldots, x_m) \\ &\underset{M}{\geq} f(x'_1, x'_2, x_3 + c_1 + c_2, \ldots, x_m) \\ &\underset{M}{\geq} \cdots \\ &\underset{M}{\geq} f\left(x'_1, \ldots, x'_m + \sum_{i=1}^{m} c_i\right) \\ &\underset{M}{\geq} f(x'_1, \ldots, x'_m). \end{aligned}$$

Conversely, suppose $\mathbf{x}' \notin \Omega(\mathbf{x})$, then $\exists i$ such that $\sum_{j=1}^{i} x_j < \sum_{j=1}^{i} x'_j$. Let $c = \sum_{j=1}^{i} x_j$. Consider the function

$$f(x_1, \ldots, x_m) = 1_{\sum_{j=1}^{i} x_j > c}.$$

Clearly, f satisfies the individual and strong pairwise monotonicity. However, we have

$$0 = f(\mathbf{x}) \underset{M}{\leq} f(\mathbf{x}') = 1.$$

Thus, we conclude.

□

A.8 Proof of Proposition 5.1

Proof. By individual monotonicity, we know if $\mathbf{x}' \in \phi(\mathbf{x})$, then $\widehat{\mu}(\mathbf{x}') \underset{M}{\leq} \widehat{\mu}(\mathbf{x})$. □

A.9 Proof of Proposition 5.2

Proof. If there is $\mathbf{x}' \in \Omega(\mathbf{x})$ has not been explored, then $\mathbf{x}' + \mathbf{e}_i$ has not been explored for all i except at boundaries. By Proposition 4.6, Theorem 4.7 and Proposition 4.11, we know $\mathbf{x}' \leq \widetilde{\mathbf{x}}$ for some $\widetilde{\mathbf{x}} \in \partial\Omega(\mathbf{x})$ and all maximum boundary points are included in the initial list. It is possible to reach max boundary points if we continue adding $\mathbf{e}_i$ for some i, as a contradiction. □

B DATA and MODELS

B.1 Finance - Credit Scoring

B.1.1 Data Description

We use the Kaggle credit score dataset [2].

- $x_1 - x_3$: Last two years, the number of times borrower was 90+ days past due, 60-89 days past due, and 30-59 days past due.
- x_4: Monthly income.
- x_5: Number of dependents in the family.
- x_6: Total balance on credit cards and personal lines of credit except for real estate and no installment debt such as car loans divided by the sum of credit limits.
- x_7: Monthly debt payments, alimony, and living costs divided by monthly gross income.
- x_8: Number of open loans and lines of credit
- x_9: Number of mortgage and real estate loans
- x_{10}: Age of borrower in years.
- y: Client's behavior; 1 = Person experienced 90 days past due delinquency or worse.

We impose strong pairwise monotonicity of $x_1 - x_3$ and individual monotonicity for $x_4 - x_5$.

For simplicity, data with missing variables are removed. Past dues that are greater or equal to 20 are discarded. Then past dues greater than four times are replaced by four due to the rarity. This also applies to x_5 if its value exceeds five. To apply the fast marching method, we categorize x_4 into the following intervals: $[0, \$2500)$, $[\$2,500, \$5,000)$, $[\$5,000, \$7,500)$, $[\$7,500, \$10,000)$, $[\$10,000, \$50,000)$, and $[\$50,000, \infty)$. Afterward, they are transformed from five to zero so that f increases monotonically with respect to x_4. We make such a choice in order to make features as easy to understand as possible for customers. This is not a unique choice. The model performance has been monitored to ensure that the accuracy does not deteriorate. When checking for accuracy, the dataset is randomly partitioned into 70% training and 30% test sets.

[2] https://www.kaggle.com/c/GiveMeSomeCredit/overview

B.1.2 Model

For MGNAM, we consider the architecture

$$f(\mathbf{x}) = f_{1,2,3}(x_1, x_2, x_3) + f_4(x_4) + \cdots + f_{10}(x_{10}). \tag{27}$$

In other words, $x_1 - x_3$ are grouped together, and the remaining features are handled using 1-dimensional functions. For $x_1 - x_3$, we enforce strong pairwise monotonicity. We enforce individual monotonicity for $x_4 - x_5$. All functions are approximated by neural networks with one hidden layer of four neurons. We focus on simple architectures since there is no apparent improvement in accuracy for more complicated models.

B.1.3 Results

The area-under-the-curve (AUC) of the model is around 85%, which indicates that the model is accurate. It might be possible to improve model performance by further cleaning the data, but since this is not the primary concern of our study, we opt to omit it for simplicity.

B.2 Criminology - Recidivism

B.2.1 Data Description

COMPAS is a proprietary score developed to predict recidivism risk, which is used to guide bail, sentencing, and parole decisions. A report published by ProPublica in 2016 provided recidivism data for defendants in Broward County, Florida (Pro, 2016). We focus on the simplified cleaned dataset provided in Dressel & Farid (2018). Three thousand and fifty-one (45%) of the 7,214 observations committed a crime within two years. This study uses a binary response variable, recidivism, as the response variable. The dataset here contains nine features selected after some feature selection was conducted.

- x_1: Total number of juvenile felony criminal charges
- x_2: Total number of juvenile misdemeanor criminal charges
- x_3: Age
- x_4: Total number of non-juvenile criminal charges
- x_5: A numeric value corresponding to the specific criminal charge
- x_6: An indicator of the degree of the charge: misdemeanor or felony
- x_7: Races include White (Caucasian), Black (African American), Hispanic, Asian, Native American, and Others
- x_8: Sex, male or female
- x_9: A numeric value between 1 and 10 corresponds to the recidivism risk score generated by COMPAS software (a small number corresponds to a low risk, and a larger number corresponds to a high risk)
- y: Whether the defendant recidivated two years after the previous charge

To avoid discrimination, we further exclude races and sexes. The COMPAS score is also excluded as it is not the focus of this study and is correlated with other features, making its interpretation more difficult. As there are too few samples, we truncate the number of juveniles exceeding five. Otherwise, if monotonicity is requested, neural network functions will become flat, which is not helpful.

B.2.2 Model

For MGNAM, we consider the architecture

$$f(\mathbf{x}) = f_{1,2}(x_1, x_2) + f_3(x_3) + \cdots + f_6(x_6). \tag{28}$$

In other words, $x_1 - x_2$ are grouped, and the remaining features are handled using 1-dimensional functions. For $x_1 - x_2$, we enforce strong pairwise monotonicity. We further impose individual monotonicity on x_4 and x_6.

B.2.3 Result

The AUC of the model is about 72%, which is consistent with the literature (Dressel & Farid, 2018).

We calculate the global feature importance by Shap in Figure 4. We take the mean value as the baseline value $\mathbf{x}'$. This result indicates that x_3, the Age, is an essential feature.

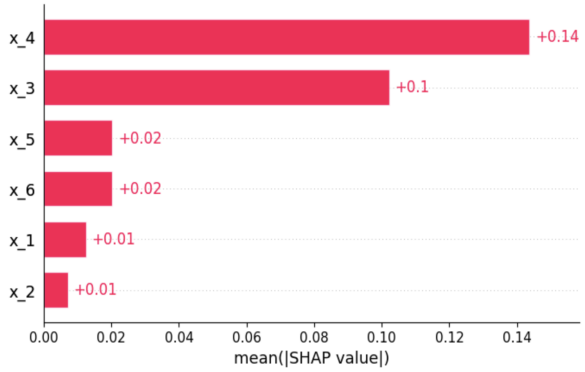


Figure 4: Global feature importance of COMPAS using Shap.

B.3 Education - Law School Bar Exam

B.3.1 Data description

The law school dataset (Wightman, 1998) concerns information on the probability of passing the bar examination. In 1991, 163 law schools in the United States were surveyed by the Law School Admission Council (LSAC). From the total of 18,692 observations, 16,856 (90%) people passed the bar for the first time. If, for instance, universities wish to award scholarships based on the likelihood of passing the bar examination, fairness could be important. In this study, the response variable is a binary variable, pass. There are 11 features in this dataset.

- x_1: The student's decile in the school given his grades in Year 3
- x_2: The student's decile in the school given his grades in Year 1
- x_3: The student's LSAT score
- x_4: The student's undergraduate GPA
- x_5: Whether the student will work full-time or part-time
- x_6: The student's family income bracket
- x_7: Tier, which is an indicator of school quality
- x_8: Whether the student is a male or female

Table 6: The comparison of the optimization 10 by considering different monotonicity for the Law school dataset with samples affected by pairwise monotonicity. Pairwise monotonicity improves the result.

	INDIVIDUAL			**WEAK**	
τ	$\frac{\lvert\mathbb{U}\rvert}{\lvert\mathbb{T}\rvert}$(%)	MEAN-ITER	τ	$\frac{\lvert\mathbb{U}\rvert}{\lvert\mathbb{T}\rvert}$(%)	MEAN-ITER
0.5	88.0	9	0.5	87.4	13
0.4	88.0	9	0.4	87.4	13
0.3	79.9	268	0.3	73.9	400
0.2	62.2	740	0.2	59.6	1220
0.1	39.5	985	0.1	38.9	1850
0	38.7	991	0	38.2	1862

- x_9: Race
- x_{10}: The student's first-year law school GPA
- x_{11}: The student's cumulative law school GPA
- y: Whether the student passed the bar exam on the first try

Race and sex were excluded for potential bias. The law school GPA (LGPA) is calculated on different scales for the first year and the cumulative. To make a comparison, we scale them. $x_{10} - x_{11}$ are excluded as they are highly correlated with $x_1 - x_2$. Additionally, to avoid unfairness, gender and race are also excluded. Hence, the first 7 features remained to train the model.

B.3.2 Model

For MGNAM, we consider the architecture

$$f(\mathbf{x}) = f_1(x_1) + f_2(x_2) + \cdots + f_7(x_7). \tag{29}$$

For all grade-related features ($x_1 - x_4$), we require individual monotonicity, as well as weak pairwise monotonicity for x_1 over x_2. In the latter cases, the requirement indicates the pairwise monotonicity of time: the more recent information should be regarded as more valuable.

B.3.3 Results

The AUC of the model is about 86%. Regarding the FMM results, approximately 86 percent of OOD data obtained a confident lower bound. The global Shap value is calculated in Figure 5. There is great significance to the x_1, x_2, and x_7 features. This model is designed to ensure fairness by not considering x_7, the tier of the law school, as a monotonic feature. However, the tier is an important feature that contributes to the uncertainty associated with the prediction. Taking the example of Figure 6, high variance data can be observed in cases where the law school's tier is high and the student's LSAT is low. Since most admitted students to top-tier schools possess high LSAT scores, this is intuitively reasonable. Without the monotonic information of tier, high-variance data cannot be effectively handled. To mitigate bias, monotonicity on tier should be avoided. If the feature tier could be replaced with other unbiased yet indicative monotonic features, such as historical bar exam passing rates of schools, we believe our performance could be further improved.

A comparison of individual monotonicity and pairwise monotonicity is presented in Table 6. The improvement has been observed.

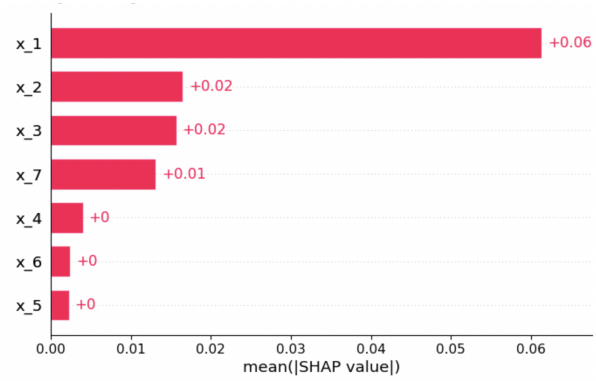


Figure 5: Global feature importance of LawSchool using Shap.

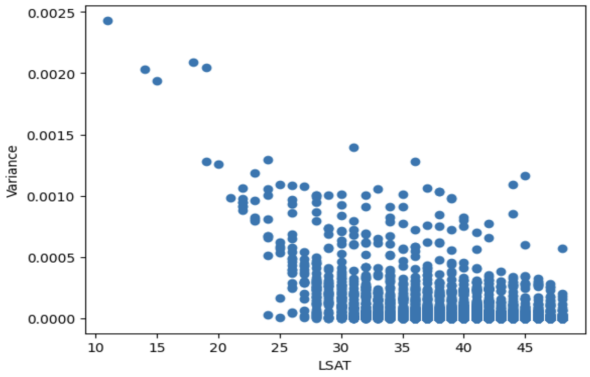


Figure 6: Variance vs LSAT in the Law School dataset.

B.4 Life science - Happiness survey

This dataset [3] is about the Somerville happiness survey by Koczkodaj (2018). The results of this study are used by the City to make decisions regarding the future development of Somerville. OOD datasets are commonly found in such surveys due to the fact that a large number of participants is unlikely. Moreover, our framework is well suited for this type of application since most of the features exhibit monotonic behavior based on domain knowledge. The following features are described.

B.4.1 Data description

- x_1: The accessibility of information regarding city services
- x_2: The housing cost
- x_3: The overall quality of public schools
- x_4: People's confidence in the local police
- x_5: Upkeep of streets and sidewalks
- x_6: Presence of social community events
- y: Whether a person is happy or not

[3] https://archive.ics.uci.edu/dataset/479/somerville+happiness+survey

B.4.2 Model

For MGNAM, we consider the architecture

$$f(\mathbf{x}) = f_1(x_1) + f_2(x_2) + \cdots + f_6(x_6). \tag{30}$$

We impose individual monotonicity on all features.

B.4.3 Results

Our model has an AUC of approximately 68 percent. The prediction of happiness is noisy by nature, so we consider the result to be satisfactory. The performance of FMM is excellent since all features are monotonic.

B.5 Medical - Mammographic Mass

B.5.1 Data description

As a screening tool for breast cancer, mammography is widely used in the medical field. However, due to the uncertainty prediction, biopsies that proved to be benign are not examined as thoroughly as they should be. There has been some previous research relating to semi-supervised learning to reduce the uncertainty of a model by Calderon-Ramirez et al. (2021). On the other hand, our approach reduces the uncertainty of OOD data by finding a lower bound, which is based on the monotonicity property. The data for Mammographic Mass is collected by Elter (2007), for a 5-feature-based classification. Overall, 961 data are available, including 516 benign and 445 malignant. Below are illustrations of all features.

- x_1: BI-RADS assessment, which is a standard assessment used by doctors to describe mammograms. The values range from 1, the benign, to 5, with a high possibility of malignancy.
- x_2: Age
- x_3: Shape of Mammography, classified into four types: round, oval, lobular, and irregular
- x_4: Margin of Mammography, classified into circumscribed, microlobulated, obscured, ill-defined, and spiculated
- x_5: Mammographic density, classified as high, iso, low and fat-containing
- y: The binary label, malignancy=1, benign=0

B.5.2 Model

For MGNAM, we consider the architecture

$$f(\mathbf{x}) = f_1(x_1) + f_2(x_2) + \cdots + f_5(x_5). \tag{31}$$

According to the doctor's diagonalization, x_1 is monotonic for predicting the severity of breast cancer. Based on previous research, there is a highly positive relationship between mammographic density and cancer severity, as discussed in Nazari & Mukherjee (2018), therefore, we impose individual monotonicity on both x_1 and x_5.

B.5.3 Results

The AUC of the model is about 90%. The OOD data is more than half due to the rare samples. If there is more data available in the hospital, this could be reduced. Based on FMM, we obtain lower bounds of confidence of approximately 30%, demonstrating the effectiveness of our approach. In the case of healthcare datasets, we believe that it is definitely possible to improve the performance by using more features as well as domain knowledge. Due to the fact that we are not experts in this field, we do not feel comfortable imposing domain knowledge in more complex situations. This is only a simple example provided for demonstration purposes.

C Algorithm Details

There are two steps in the training process for the model.

Initially, a mini-batch gradient descent approach with a batch size of 64 is applied to pre-train the model without taking into account monotonicity. The initial learning rate is 1×10^{-2}, and it is multiplied by 0.1 when there is no improvement in the loss over 5 epochs, and early stopping is performed when there is no improvement over 10 epochs.

As part of Step 2, the model is trained to satisfy all monotonicity requirements and the batch gradient descent method is applied. The α factor, which represents the punishment for features exhibiting monotonicity violations, is initially set at 1×10^{-1} and is multiplied by 10 every 10 epochs. At the same time, the learning rate at is set to 1×10^{-2}. Once all monotonies have been satisfied, the remuneration will change to 1×10^{-3} for another 10 epochs of training.

D Other Discussions

D.1 Binning

By binning or discretizing, continuous features are transformed into discrete ones. Binning is a common practice in numerical partial differential equations (Sethian, 1996). For ML, binning may improve predictive models' accuracy by reducing noise or nonlinearity in the dataset as well as identifying outliers, and invalid and missing values of numerical features. The use of bins is particularly popular in high-risk sectors, where interpretation is of paramount importance. When considering the income feature, for example, people are more likely to consider low-, median-, and high-income classes rather than specific figures. Various binning methods are available, including equal width, equal frequency, and weight of evidence. As the choice of methods depends heavily on the application and the appetite of the user, we will not discuss this further. It is possible to preserve continuous features by leaving original features alone and binning only the features in the optimization process.

D.2 Choice of Models

We apply accurate and transparent monotonic groves of the neural additive model (MGNAM) proposed in Chen & Ye (2023), as three types of monotonicity are included. The code is modified based on the Neural Additive Models (Agarwal et al., 2021). In general, the choice is not unique. Models developed by Liu et al. (2020); Milani Fard et al. (2016); You et al. (2017); Runje & Shankaranarayana (2023) are applicable for individual monotonicity, whereas deep lattice models (Gupta et al., 2020; Cotter et al., 2019) include strong pairwise monotonicity.

D.3 Choice of Threshold ϵ

ϵ thresholds are not unique and depend on applications and user preferences. In high-risk sectors, one may choose a very small value for ϵ. However, if the risks are tolerable, a larger ϵ may be chosen. The finance sector, for instance, has seen different types of investors, including risk-averse, risk-neutral, and risk-seeking investors. There is a discussion in Section 3.6 by Petters & Dong (2016).