Towards Modeling Uncertainties of Self-Explaining Neural Networks via Conformal Prediction

Wei Qian[*1], **Chenxu Zhao**[*1], **Yangyi Li**[*1], **Fenglong Ma**[2], **Chao Zhang**[3], **Mengdi Huai**[1]

[1]Iowa State University
[2]Pennsylvania State University
[3]Georgia Institute of Technology
{wqi, cxzhao, liyangyi, mdhuai}@iastate.edu, fenglong@psu.edu, chaozhang@gatech.edu

Abstract

Despite the recent progress in deep neural networks (DNNs), it remains challenging to explain the predictions made by DNNs. Existing explanation methods for DNNs mainly focus on post-hoc explanations where another explanatory model is employed to provide explanations. The fact that post-hoc methods can fail to reveal the actual original reasoning process of DNNs raises the need to build DNNs with built-in interpretability. Motivated by this, many self-explaining neural networks have been proposed to generate not only accurate predictions but also clear and intuitive insights into why a particular decision was made. However, existing self-explaining networks are limited in providing distribution-free uncertainty quantification for the two simultaneously generated prediction outcomes (i.e., a sample's final prediction and its corresponding explanations for interpreting that prediction). Importantly, they also fail to establish a connection between the confidence values assigned to the generated explanations in the interpretation layer and those allocated to the final predictions in the ultimate prediction layer. To tackle the aforementioned challenges, in this paper, we design a novel uncertainty modeling framework for self-explaining networks, which not only demonstrates strong distribution-free uncertainty modeling performance for the generated explanations in the interpretation layer but also excels in producing efficient and effective prediction sets for the final predictions based on the informative high-level basis explanations. We perform the theoretical analysis for the proposed framework. Extensive experimental evaluation demonstrates the effectiveness of the proposed uncertainty framework.

Introduction

Recent advancements in DNNs have undeniably led to remarkable improvements in accuracy. However, this progress has come at the expense of losing visibility into their internal workings (Nam et al. 2020). This lack of transparency often renders these deep learning models as "black boxes" (Li et al. 2020; Bang et al. 2021; Zhao et al. 2023). Without understanding the rationales behind the predictions, these black-box models cannot be fully trusted and widely applied in critical areas. In addition, explanations can facilitate model debugging and error analysis. These indicate the necessity of investigating the explainability of DNNs.

To offer valuable insights into how a model makes predictions, many approaches have surfaced to explain the behavior of black-box neural networks. These methods primarily focus on providing post-hoc explanations, attempting to provide insights into the decision-making process of the underlying model (Lundberg and Lee 2017; Ribeiro, Singh, and Guestrin 2016; Slack et al. 2020, 2021; Yao et al. 2022; Han, Srinivas, and Lakkaraju 2022; Huai et al. 2020). Note that these post-hoc methods need designing another explanatory model to provide explanations for a trained deep learning model. Despite their widespread use, post-hoc interpretation methods can be inaccurate or incomplete in revealing the actual reasoning process of the original model (Rudin 2018).

To address these challenges, researchers have been actively promoting the adoption of inherently self-explaining neural networks (Koh et al. 2020; Havasi, Parbhoo, and Doshi-Velez 2022; Chauhan et al. 2023; Yuksekgonul, Wang, and Zou 2022; Alvarez Melis and Jaakkola 2018; Sinha et al. 2023), which integrate interpretability directly into the model architecture and training process, producing explanations that are utilized during classification. Based on the presence of ground-truth basis concepts during training, existing works on self-explaining networks can be broadly classified into two groups: those with access to ground-truth and those without. For example, (Koh et al. 2020) trains the concept bottleneck layer to align concept predictions with established true concepts, while (Alvarez Melis and Jaakkola 2018) utilizes autoencoder techniques to learn prototype-based concepts from the training data.

Although existing self-explaining networks can offer inherent interpretability to make their decision-making process transparent and accessible, they are limited in the distribution-free uncertainty quantification for both the generated explanation outcomes and the corresponding label predictions. Such limitations have the potential to significantly impede the widespread acceptance and utilization of self-explaining neural networks. In practice, such confidence values serve as indicators of the likelihood of correctness for each prediction, and understanding the likelihood of each prediction allows us to assess the degree to which we can rely on it. Note that self-explaining neural networks simultaneously generate two outputs (i.e., the final prediction decision and an explanation of that final decision). The most straightforward way is to separately and

independently quantify uncertainties for the predicted explanations and their corresponding final predictions. Nevertheless, we argue that within self-explaining neural networks, the high-level interpretation space is informative due to its powerful inductive bias in generating high-level basis concepts for interpretable representations. Imagine a scenario in which we engage in the binary classification task distinguishing between zebras and elephants, and we hold a strong 95% confidence in identifying the presence of the "Stripes" concept within a test image sample. In this context, we can assert that the model exhibits high confidence in assigning the "Zebra" label to this particular test sample. This also aligns with existing class-wise explanations (Li et al. 2021; Zhao et al. 2021) that focus on uncovering the informative patterns that influence the model's decision-making process when classifying instances into a particular class category.

Our goal in this paper is to offer distribution-free uncertainty quantification for self-explaining networks via two indispensable and interconnected aspects: establishing a reliable uncertainty measure for the generated explanations and effectively transferring the explanation prediction sets to the associated predicted class outcomes. Note that the recent work of (Slack et al. 2021) focuses exclusively on post-hoc explanations and relies on certain assumptions regarding the properties of these explanations, such as assuming a Gaussian distribution of explanation errors. On the other hand, (Kim et al. 2023) mainly presents probabilities and uncertainties for individual concepts or classes, and does not provide the prediction set with uncertainty, neither for the concept layer nor for the prediction layer. Therefore, in addition to providing distribution-free uncertainty quantification for the generated basis explanations, it is also essential to study how to transfer the constructed prediction sets in the interpretation space to the final prediction decision space.

To address the above challenges, in this paper, we design a novel **un**certainty modeling framework for **s**elf-**e**xplaining **n**eural **n**etworks (**unSENN**), which can make valid and reliable predictions or estimates with quantifiable uncertainty, even in situations where the underlying data distribution may not be well-defined or where traditional assumptions of parametric statistics might not hold. Specifically, in our proposed method, we first design novel non-conformity measures to capture the informative high-level concepts in the interpretation layer for self-explaining neural networks. Importantly, our proposed non-conformity measures can overcome the difficulties posed by self-explaining networks, even in the absence of explanation ground truth in the interpretation layer. Note that these calculated scores reflect the prediction error of the underlying explainer, where a smaller prediction error would lead to the construction of smaller and more informative concept sets. We also prove the theoretical guarantees for the constructed concept prediction sets in the interpretation layer. To get better conformal predictions for the final predicted labels, we then design effective transfer functions, which involve the search of possible labels under certain constraints. Due to the challenge posed by directly estimating final prediction sets through the formulated transfer functions, we transform them into new optimization frameworks from the perspective of adversarial attacks. We also perform the theoretical analysis for the final prediction sets. The extensive experimental results verify the effectiveness of our proposed method.

Problem Formulation

Let us consider a general multi-class classification problem with a total of $M > 2$ possible labels. For an input $x \in \mathbb{R}^d$, its true label is represented as $y \in \mathbb{R}$. Note that self-explaining networks generate high-level concepts in the interpretation layer prior to the output layer. In Definition 1, we give a general definition of a self-explaining model (Alvarez Melis and Jaakkola 2018; Elbaghdadi et al. 2020).

Definition 1 (Self-explaining Models). *Let $x \in \mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}^M$ be the input and output spaces. We say that $f : \mathcal{X} \to \mathcal{Y}$ is a self-explaining prediction model if it has the following form*

$$f(x) = g(\theta_1(x)h_1(x), \cdots, \theta_C(x)h_C(x)), \quad (1)$$

where the prediction head g depends on its arguments, the set $\{(\theta_c(x), h_c(x))\}_{c=1}^{C}$ consists of basis concepts and their influence scores, and C is small compared to d.

Example 1: Concept bottleneck models (Koh et al. 2020). Concept bottleneck models (CBMs) are interpretable neural networks that first predict labels for human-interpretable concepts relevant to the prediction task, and then predict the final label based on the concept label predictions. We denote raw input features by x, the labels by y, and the pre-defined true concepts by $c \in \mathbb{R}^C$. Given the observed training samples $D^{tra} = \{(x_i, c_i, y_i)\}_{i=1}^{N^{tra}}$, a CBM (Koh et al. 2020) can be trained using the concept and class prediction losses via two distinct approaches: *independent* and *joint*.

Example 2: Prototype-based self-explaining networks (Alvarez Melis and Jaakkola 2018). Different from CBMs (Koh et al. 2020), prototype-based self-explaining networks usually use an autoencoder to learn a set of prototype-based concepts directly from the training data (i.e., $D^{tra} = \{(x_i, y_i)\}_{i=1}^{N^{tra}}$) during the training process without the pre-defined concept supervision information. These prototype-based concepts are extracted in a way that they can best represent specific target sets (Zhang, Gao, and Ma 2023).

For a well-trained self-explaining network $f = g \circ h$, our goal in this paper is to quantify the uncertainty of the generated explanations in the interpretation layer (preceding the output layer) and the final predictions in the output layer. In particular, for a test sample x and a chosen miscoverage rate $\varepsilon \in (0, 1)$, our primary objective is twofold: First, we aim to calculate a set $\Gamma_{cpt}^{\varepsilon}(x)$ containing x's true concepts with probability at least $1 - \varepsilon$. In addition, we will investigate how to transfer the explanation prediction sets to the final decision sets. In practice, predictions associated with confidence values are highly desirable in risk-sensitive applications.

Methodology

In this section, we take concept bottleneck models (Koh et al. 2020; Havasi, Parbhoo, and Doshi-Velez 2022; Chauhan et al. 2023; Yuksekgonul, Wang, and Zou 2022) as an illustrative example to present our proposed method. Note

that concept bottleneck models incorporate pre-defined concepts into a supervised learning procedure. We denote the available dataset as $D = \{(x_i, c_i, y_i)\}_{i=1}^N$, where $x_i \in \mathbb{R}^d$, $c_i = [c_{i,1}, \cdots, c_{i,C}] \in \mathbb{R}^C$ and $y_i \in [M]$. These samples are drawn exchangeably from some unknown distribution P_{XCY}. To train a self-explaining network, we first split the available dataset $D = \{(x_i, c_i, y_i)\}_{i=1}^N$ into a training set D^{tra} and a calibration set D^{cal}, where $D^{tra} \cap D^{cal} = \emptyset$ and $D^{tra} \cup D^{cal} = D$. Given the available training set D^{tra} associated with the supervised concepts, we can train a self-explaining network $f = g \circ h$, where $h : \mathbb{R}^d \to \mathbb{R}^C$ maps a raw sample x into the concept space and $g : \mathbb{R}^C \to \mathbb{R}^M$ maps concepts into a final class prediction.

Given the underlying well-trained self-explaining model $f = g \circ h$, our goal is to provide validity guarantees on its explanation and final prediction outcomes in a *post-hoc* way. Note that each x_i, its true relevant concepts are represented by a concept vector $c_i = [c_{i,1}, \cdots, c_{i,j}, \cdots, c_{i,C}] \in \{0,1\}^C$, with $c_{i,j} = 1$ indicating that x_i is associated with the j-th concept. We also use $\mathcal{C}_{cpt}(x_i) = \{j | c_{i,j} = 1\}$ to represent the set of true concepts for x_i. For x_i, we can use $h : \mathbb{R}^d \to \mathbb{R}^C$ to obtain $h(x_i) = [h_1(x_i), \cdots, h_j(x_i), \cdots, h_C(x_i)]$, where $h_j(x_i)$ is the prediction score of x_i with regards to the j-th concept. We denote $[h_{[1]}(x_i), \cdots, h_{[j]}(x_i), \cdots, h_{[C]}(x_i)]$ as the sorted values of $h(x_i)$ in the descending order, i.e., $h_{[1]}(x_i)$ is the largest (top-1) score, $h_{[2]}(x_i)$ is the second largest (top-2) score, and so on. Furthermore, $[j]$ corresponds to the concept index of the top-j concept prediction score, i.e., $j' = [j]$ if $h_{j'}(x_i) = h_{[j]}(x_i)$. In practice, researchers are usually interested in identifying the top-K important concepts to help users understand the key factors driving the model's predictions (Agarwal et al. 2022; Brunet, Anderson, and Zemel 2022; Yeh et al. 2020; Huai et al. 2022; Rajagopal et al. 2021; Hitzler and Sarker 2022). Here, we can easily convert a general concept predictor h into a top-K concept classifier by returning the set of concepts corresponding to the set of top-K prediction scores from h. For x_i, the top-K concept prediction function returns the set $\tilde{\mathcal{C}}_{cpt}(x_i) = \{[1], \cdots, [K]\}$ for $1 \leq K < C$. Here, we assume the concept prediction scores are calibrated, i.e., taking values in the range of $[0,1]$. For $\Psi \in \mathbb{R}$, we can use simple transforms such as $\sigma(\Psi) = 1/(1 + e^{-\Psi})$ to map it to the range of $[0,1]$ without changing their ranking. Then, for x_i, we obtain the below calibrated concept importance scores

$$\tilde{h}(x_i) = [1/(1 + \exp(-h_1(x_i))), \cdots, 1/(1 \quad\quad (2)$$
$$+ \exp(-h_C(x_i)))] = [\sigma(h_1(x_i)), \cdots, \sigma(h_C(x_i))],$$

where $\sigma(h_j(x_i)) \in [0,1]$ is the calibrated importance score of x_i with regards to the j-th concept. Not that for x_i, it is associated with a concept subset $\mathcal{C}_{cpt}(x_i)$, which can be represented by a vector $h^*(x_i) = [h_1^*(x_i), \cdots, h_j^*(x_i), \cdots, h_C^*(x_i)]$, where $h_j^*(x_i) = 1$ if and only if concept j is associated with x_i, and 0 otherwise.

In order to quantify how "strange" the generated concept explanations are for a given sample x_i, we utilize the underlying well-trained self-explaining model f to construct the following non-conformity measure

$$s(x_i, h^*(x_i), \tilde{h}(x_i)) = \sum_{j=1}^K (1 - 1/(1 + \exp(-h_{[j]}(x_i))))$$

$$+ \lambda_2 * \sum_{j=K+1}^C (1/(1 + \exp(-h_{[j]}(x_i)))) = \sum_{j=1}^K |h_{[j]}^*(x_i)$$

$$- \tilde{h}_{[j]}(x_i)| + \lambda_2 * \sum_{j=K+1}^C |h_{[j]}^*(x_i) - \tilde{h}_{[j]}(x_i)|, \quad (3)$$

where λ_2 is a pre-defined trade-off parameter, and $[j]$ corresponds to the concept index of the top-j concept prediction score of $\tilde{h}(x_i)$. The above non-conformity measure returns small values when the concept prediction outputs of the true relevant concepts are high and the concept outputs of the non-relevant concepts are low. Note that smaller conformal scores correspond to more model confidence (Angelopoulos and Bates 2021; Teng, Tan, and Yuan 2021). In particular, when $\lambda_2 = 0$, we only consider the top-ranked concepts.

Based on the above constructed non-conformity measure, for each calibration sample $(x_i, c_i, y_i) \in D^{cal}$, we apply the non-conformity measure $s(x_i, h^*(x_i), \tilde{h}(x_i))$ to get N^{cal} non-conformity scores $\{s_i\}_{i=1}^{N^{cal}}$, where N^{cal} is the number of calibration samples. Then, we select a miscoverage rate $\varepsilon \in (0,1)$ and compute $Q_{1-\varepsilon}$ as the $\frac{\lceil (N^{cal}+1)(1-\varepsilon) \rceil}{N^{cal}}$ quantile of the non-conformity scores $\{s_i\}_{i=1}^{N^{cal}}$, where $\lceil \cdot \rceil$ is the ceiling function. Formally, $Q_{1-\varepsilon}$ is defined as

$$Q_{1-\varepsilon} = \inf\{Q : \frac{|\{i : s(x_i, h^*(x_i), \tilde{h}(x_i)) \leq Q\}|}{N^{cal}} \quad (4)$$
$$\geq \frac{\lceil (N^{cal} + 1)(1 - \varepsilon) \rceil}{N^{cal}}\}.$$

Based on the above, given $Q_{1-\varepsilon}$ and a desired coverage level $1 - \varepsilon \in (0,1)$, for a test sample x^{test} (where x^{test} is known but the true top-K concepts $\mathcal{C}$ are not), we can get the explanation prediction set $\Gamma_{cpt}^\varepsilon(x^{test})$ for x^{test} with $1 - \varepsilon$ coverage guarantee. Algorithm 1 gives the procedure to calculate the concept prediction sets.

Definition 2 (Data Exchangeability). *Consider variables $z_1, \cdots, z_N$. The variables are exchangeable if for every permutation τ of the integers $1, \cdots, N$, the variables $w_1, \cdots, w_N$ where $w_i = z_{\tau(i)}$, have the same joint probability distribution as $z_1, \cdots, z_N$.*

Theorem 1. *Suppose that the calibration samples ($D^{cal} = \{(x_i, c_i, y_i)\}_{i=1}^{N^{cal}}$) and the given test sample x^{test} are exchangeable. Then, if we calculate the quantile value $Q_{1-\varepsilon}$ and construct $\Gamma_{cpt}^\varepsilon(x^{test})$ as indicated above, for the above non-conformity score function s and any $\varepsilon \in (0,1)$, the derived $\Gamma_{cpt}^\varepsilon(x^{test})$ satisfies*

$$P(\mathcal{C}_{cpt}(x^{test}) \in \Gamma_{cpt}^\varepsilon(x^{test})) \geq 1 - \varepsilon, \quad (5)$$

where $\mathcal{C}_{cpt}(x^{test})$ is the true relevant concept set for the given test sample x^{test}, and ε is a desired miscoverage rate.

Algorithm 1: Uncertainty quantification for self-explaining neural networks

---

1: **Input:** Calibration data D^{cal}, pre-trained model $f = g \circ h$, non-conformity measure $s(\cdot)$, a test sample x^{test}, and a desired miscoverage rate $\varepsilon \in (0, 1)$.
2: **Output:** The concept prediction set $\Gamma^\varepsilon_{cpt}(x^{test})$ and label set $\Gamma^\varepsilon_{lab}(x^{test})$ for sample x^{test}.
3: Calculate the non-conformity scores on the calibration samples (i.e., D^{cal}) based on Eqn. (3)
4: Compute the quantile value $Q_{1-\varepsilon}$ based on Eqn. (4)
5: Construct the concept prediction set $\Gamma^\varepsilon_{cpt}(x^{test})$ for x^{test} with $1 - \varepsilon$ coverage guarantee
6: Initialize $\Gamma^\varepsilon_{lab}(x^{test}) = \emptyset$
7: **for** $m = 1, 2, \cdots, M$ **do**
8:    Obtain δ by optimizing the loss in Eqn. (10)
9:    **if** $\sum_{m' \in [M] \setminus m} \mathbb{I}[g_m(v) > g_{m'}(v)] == M - 1$ **then**
10:       Update $\Gamma^\varepsilon_{lab}(x^{test}) = \Gamma^\varepsilon_{lab}(x^{test}) \cup \{m\}$
11:    **end if**
12: **end for**
13: **Return:** $\Gamma^\varepsilon_{cpt}(x^{test})$ and $\Gamma^\varepsilon_{lab}(x^{test})$.

Proof. Let $s_1, s_2, \cdots, s_{N^{cal}}, s_{test}$ denote the non-formality scores of the calibration samples and the given test samples. To avoid handling ties, we consider the case where these non-formality scores are distinct with probability 1. Without loss of generality, we assume these calibration scores are sorted so that $s_1 < s_2 < \cdots < s_{N^{cal}}$. In this case, we have that $Q_{1-\varepsilon} = s_{\lceil (N^{cal}+1)(1-\varepsilon) \rceil}$ when $\varepsilon \geq \frac{1}{N^{cal}+1}$ and $Q_{1-\varepsilon} = \infty$ otherwise. Note that in the case where $Q_{1-\varepsilon} = \infty$, the coverage property is trivially satisfied; thus, we only need to consider the case when $\varepsilon \geq \frac{1}{N^{cal}+1}$. We begin by observing the below equality of the two events

$$\{\mathcal{C}_{cpt}(x^{test}) \in \Gamma^\varepsilon_{cpt}(x^{test})\} = \{s_{test} \leq Q_{1-\varepsilon}\}. \quad (6)$$

By combining this with the definition of $Q_{1-\varepsilon}$, we have

$$\{\mathcal{C}_{cpt}(x^{test}) \in \Gamma^\varepsilon_{cpt}(x^{test})\} = \{s_{test} \leq s_{\lceil (N^{cal}+1)(1-\varepsilon) \rceil}\}.$$

Since the variables $s_1, \cdots, s_{test}$ are exchangeable, we can have that $P(s_{test} \leq s_k) = k/(N^{cal} + 1)$ holds true for any integer k. This means that s_{test} is equally likely to fall in anywhere between the calibration scores $s_1, \cdots, s_{N^{cal}}$. Notably, the randomness is over all variables. Building on this, we can derive that $P(s_{test} \leq s_{\lceil (N^{cal}+1)(1-\varepsilon) \rceil}) = \frac{\lceil (N^{cal}+1)(1-\varepsilon) \rceil}{(N^{cal}+1)} \geq 1-\varepsilon$, which yields the desired result. $\square$

In the above theorem, we show that the proposed algorithm can output predictive concept sets that are rigorously guaranteed to satisfy the desired coverage property shown in Eqn. (5), no matter what (possibly incorrect) self-explaining model is used or what the (unknown) distribution of the data is. The above theorem relies on the assumption that the data points are exchangeable (see Definition 2), which is widely adopted by existing conformal inference works (Cherubin, Chatzikokolakis, and Jaggi 2021; Tibshirani et al. 2019). The data exchangeability assumption is much weaker than i.i.d. Note that i.i.d variables are automatically exchangeable

(Sesia and Favaro 2022). For a random variable $v \in \mathbb{R}^C$, we denote $[v_{[1]}, v_{[2]}, \cdots, v_{[C]}]$ as the sorted values of v in the descending order, i.e., $v_{[1]}$ is the largest (top-1) score, $v_{[2]}$ is the second largest (top-2) score, and so on. Then, we can construct the label prediction set for the test sample x^{test} as

$$\Gamma^\varepsilon_{lab}(x^{test}) = \{y = g(v) : \sum_{j=1}^{K} |h^*_{[j]}(x^{test}) - \tilde{h}_{[j]}(v)|$$

$$+ \lambda_2 * \sum_{j=K+1}^{C} |h^*_{[j]}(x^{test}) - \tilde{h}_{[j]}(v)| \leq Q_{1-\varepsilon}\}, \quad (7)$$

where $\tilde{h}(v) = \sigma(v)$. In the above, since the true concepts are unknown for x^{test}, we instead calculate $h^*_{[j]}(x^{test}) = \tilde{h}_{[j]}(x^{test}) = \sigma(h_{[j]}(x^{test}))$. Without loss of generality, in the following, we take a setting where $\lambda_2 = 1$ as an example to discuss how to transfer the confidence prediction sets in the concept space to the final class label prediction layer, and the *extension* to other cases can be easily derived. In this case, we calculate the set for x^{test} as $\Gamma^\varepsilon_{lab}(x^{test}) = \{y = g(v) : ||\tilde{h}(v) - h^*(x^{test})||_1 \leq Q_{1-\varepsilon}\}$, where $v \in \mathbb{R}^C$, $\hat{v} = \sigma(v) = [\sigma(v_1), \cdots, \sigma(v_C)]$, g is the prediction head, $\tilde{h}(x^{test}) = [\sigma(h_1(x^{test})), \cdots, \sigma(h_C(x^{test}))]$ and $Q_{1-\varepsilon}$ is derived based on the calibration set. To derive the label sets, we propose to solve the following optimization

$$\max_{v \in \mathbb{R}^C} \sum_{m' \in [M] \setminus m} \mathbb{I}[g_m(v) > g_{m'}(v)] \quad (8)$$

$$s.t., ||\tilde{h}(v) - h^*(x^{test})||_1 \leq Q_{1-\varepsilon},$$

where the quantile value $Q_{1-\varepsilon}$ is derived based on Eqn. (4). In the above, we want to find a v, whose predicted label is "m", under the constraint that $||\tilde{h}(v) - h^*(x^{test})||_1 \leq Q_{1-\varepsilon}$. If we can successfully find such a v, we will include the label "m" in the label set; otherwise, we will exclude it. Notably, we need to iteratively solve the above optimization over all possible labels to construct the final prediction set.

However, it is difficult to directly optimize the above problem, due to the non-differentiability of the discrete components in the above equation. Below, we use $\delta = [\delta_1, \cdots, \delta_C]$ to denote the difference between $\hat{v}$ and $\tilde{h}(x^{test})$, i.e., $\delta = \hat{v} - \tilde{h}(x^{test})$, where $\hat{v} = \sigma(v) = [\sigma(v_1), \cdots, \sigma(v_C)]$, and $\tilde{h}(x^{test}) = \sigma(h(x^{test}))$. Based on this, we can have

$$v = [v_1 = -\log(1/(\tilde{h}_1(x^{test}) + \delta_1) - 1), \cdots, \quad (9)$$

$$v_C = -\log(1/(\tilde{h}_C(x^{test}) + \delta_C) - 1)].$$

To address the above challenge, we propose to solve the below reformulated optimization

$$\min_{||\delta||_1 \leq Q_{1-\varepsilon}} \max(\max_{m' \neq m} g_{m'}(v) - g_m(v), -\beta), \quad (10)$$

where $m' \in [M]$, β is a pre-defined value, v is derived based on Eqn. (9), and $g : \mathbb{R}^C \to \mathbb{R}^M$ maps v into a final class prediction. $g(v)$ represents the logit output of the underlying self-explaining model for each class when the input is v. The above adversarial loss enforces the actual prediction of

v to the targeted label "m" (Huang et al. 2020; Carlini and Wagner 2017). If we can find such a v, the label "m" will be included in the prediction set; otherwise, it will not.

Lemma 2 ((Tibshirani et al. 2019)). *If $V_1, \cdots, V_{n+1}$ are exchangeable random variables, then for any $\gamma \in (0, 1)$, we have the following*

$$P\{V_{n+1} \leq Quantile(\gamma; V_{1:n} \cup \{\infty\})\} \geq \gamma, \qquad (11)$$

where $Quantile(\gamma; V_{1:n} \cup \{\infty\})$ denotes the level γ quantile of $V_{1:n} \cup \{\infty\}$. In the above, $V_{1:n} = \{V_1, \cdots, V_n\}$.

Theorem 3. *Suppose the calibration samples and the test sample are exchangeable, if we construct $\Gamma_{lab}^{\varepsilon}(x^{test})$ as indicated above, the following inequality holds for any $\varepsilon \in (0, 1)$*

$$P(y^{test} \in \Gamma_{lab}^{\varepsilon}(x^{test})) \geq 1 - \varepsilon, \qquad (12)$$

where y^{test} is the true label of x^{test}, and $\Gamma_{lab}^{\varepsilon}(x^{test})$ is the label prediction set derived based on Eqn. (7).

Due to space constraints, the complete proof for Theorem 3 is postponed to the full version of this paper. This proof is based on Lemma 2.

Discussions on prototype-based self-explaining neural networks. Here, we will discuss how to generalize our above method to the prototype-based self-explaining networks, which do not depend on the pre-defined expert concepts (Alvarez Melis and Jaakkola 2018; Li et al. 2018). They usually adopt the autoencoder network to learn a lower-dimension latent representation of the data with an encoder network, and then use several network layers over the latent space to learn C prototype-based concepts, i.e., $\{p_j \in \mathbb{R}^q\}_{j=1}^C$. After that, for each p_j, the similarity layer computes its distance from the learned latent representation (i.e., $e_{enc}(x)$) as $h_j(x) = ||e_{enc}(x) - p_j||_2^2$. The smaller the distance value is, the more similar $e_{enc}(x)$ and the j-th prototype (p_j) are. Finally, we can output a probability distribution over the M classes. Nonetheless, we have no access to the true basis explanations if we want to conduct conformal predictions at the prototype-based interpretation level. To address this, for prototype-based self-explaining networks, we construct the following non-conformity measure

$$s(x_i, y_i, f) = \inf_{v \in \{v: \ell(g(v), y_i) \leq \alpha\}} [\sum_{j=1}^{K} |h_{[j]}(x_i) - v_{[j]}|$$

$$+ \lambda_2 * \sum_{j=K+1}^{C} |h_{[j]}(x_i) - v_{[j]}|], \qquad (13)$$

where $\ell(g(v), y_i)$ is the classification cross entropy loss for v, and α is a predefined small threshold value to make the prediction based on v to be close to label y_i. Based on the above non-conformity measure, we can calculate the non-conformity scores for the calibration samples, and then calculate the desired quantile value. Based on this, for the given test sample, following Eqn. (8) and (13), we can calculate the final prediction set. However, it is usually complicated to calculate the above score due to the infimum operator. To overcome this challenge, we apply gradient descent starting
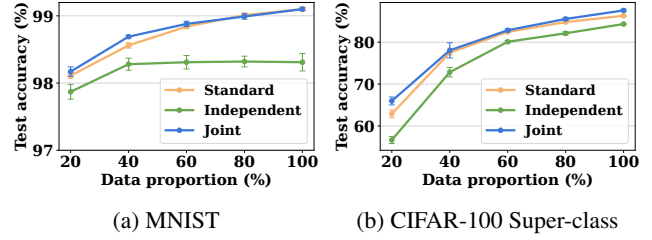

Figure 1: Data efficiency for concept bottleneck models with different learning strategies.

from the trained similarity vector to find a surrogate vector around it. Additionally, to obtain the final prediction sets, we follow Eqn. (10) to determine whether a specific class label should be included in the final prediction sets.

Experiments

In this section, we conduct experiments to evaluate the performance of the proposed method (i.e., unSENN). Due to space limitations, more experimental details and results (e.g., experiments on more self-explaining models and running time) can be found in the full version of this paper.

Experimental Setup

Real-world datasets. In experiments, we adopt the following real-world datasets: **CIFAR-100 Super-class** (Fischer et al. 2019) and **MNIST** (Deng 2012). Specifically, the CIFAR-100 Super-class dataset comprises 100 diverse classes of images, which are further categorized into 20 super-classes. For example, the five classes *baby, boy, girl, man* and *woman* belong to the super-class *people*. In this case, we exploit each image class as a concept-based explanation for the super-class prediction (Hong and Pavlic 2023). The MNIST dataset consists of a collection of grayscale images of handwritten digits (0-9). In experiments, we consider the MNIST range classification task, wherein the goal is to classify handwritten digits into five distinct classes based on their numerical ranges. Specifically, the digits are divided into five non-overlapping ranges (i.e., $[0, 1], [2, 3], [4, 5], [6, 7], [8, 9]$), each assigned a unique label (the label set is denoted as $\{0, 1, 2, 3, 4\}$). For example, an input image depicting a handwritten digit "4" is classified as class 2 within the range of digits 4 to 5. Here, we exploit the identity of each digit as a concept-based explanation for the ranges prediction (Rigotti et al. 2021).

Baselines. In experiments, we adopt two baselines: *Naive* baseline and *Vanilla Conformal Predictors*. These baselines exclusively rely upon $\mathcal{X}$ and $\mathcal{Y}$. Specifically, the Naive baseline constructs the set sequentially incorporating classes in decreasing order of their probabilities until the cumulative probability surpasses the predefined threshold $1 - \varepsilon$. Notably, this Naive approach lacks a theoretical guarantee of its coverage. The Vanilla Conformal Predictors utilize the final decisions to construct the prediction sets. Following (Cherubin, Chatzikokolakis, and Jaggi 2021), we set the non-conformity measure for this Vanilla baseline as
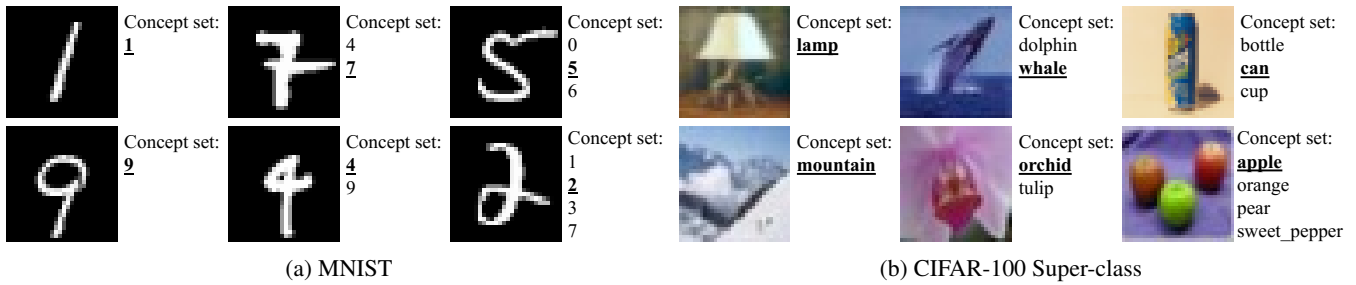
(a) MNIST   (b) CIFAR-100 Super-class

Figure 2: Concept conformal sets on MNIST and CIFAR-100 Super-class. Phrases in bold and underlined mean true concepts.

Dataset	Miscoverage rate ε	Concept error rate	
		unSENN	Naive
MNIST	0.05	0.047 ± 0.001	0.003 ± 0.000
	0.1	0.094 ± 0.002	0.003 ± 0.000
	0.15	0.139 ± 0.006	0.004 ± 0.000
	0.2	0.174 ± 0.011	0.005 ± 0.000
CIFAR-100 Super-class	0.05	0.044 ± 0.002	0.073 ± 0.001
	0.1	0.091 ± 0.003	0.100 ± 0.001
	0.15	0.137 ± 0.004	0.121 ± 0.001
	0.2	0.177 ± 0.003	0.136 ± 0.001

Table 1: Error rate of concept conformal sets on MNIST and CIFAR-100 Super-class with different miscoverage rates.

$s(x, y) = -f_y(x)$, where f is a pre-trained model, and $f_y(x)$ is the softmax probability for x with true label y.

Implementation details. In experiments, we evaluate the proposed method (i.e., unSENN) across the following underlying models: ResNet-50 (He et al. 2016), a convolutional neural network (CNN), and a multi-layer perception (MLP). For the calibration set, we randomly hold out 10% of the original available dataset to compute the non-conformity scores. All the experiments are conducted 10 times, and we report the mean and standard errors.

Experimental Results

First, we measure the data efficiency (the amount of training data needed to achieve a desired level of accuracy) of the self-explaining model. For the CIFAR-100 Super-class, we apply ResNet-50 and MLP to learn the concepts and labels, respectively. Similarly, we use CNN and MLP for the MNIST. In Figure 1, we present the test accuracy of the self-explaining model using independent and joint learning strategies with various data proportions. Note that the standard model ignores the concepts and directly maps input features to the final prediction. As we can see, the self-explaining models are particularly effective on the adopted datasets. The joint model exhibits comparable performance to the standard model, and the joint model is slightly more accurate in lower data regimes (e.g., 20%). Therefore, in the following experiments, we adopt the joint learning strategy.

Then, we explore the effectiveness of the concept conformal sets within the self-explaining model. In Table 1, we report the error rate (the percentage of true concepts that are

Dataset	Method	Label coverage	Average set size
MNIST	unSENN	0.988 ± 0.002	1.00 ± 0.00
	Naive	0.985 ± 0.002	1.01 ± 0.00
	Vanilla	0.903 ± 0.001	0.90 ± 0.00
CIFAR-100 Super-class	unSENN	0.847 ± 0.002	1.03 ± 0.00
	Naive	0.786 ± 0.004	1.28 ± 0.02
	Vanilla	0.805 ± 0.006	1.15 ± 0.01

Table 2: Label coverage and average set size of prediction conformal sets on MNIST and CIFAR-100 Super-class.

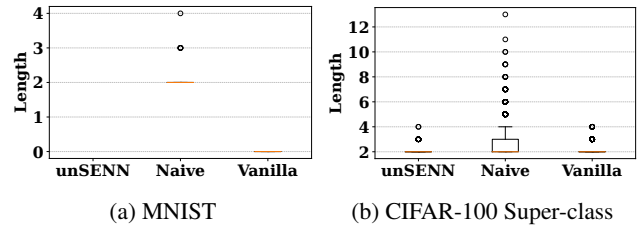(a) MNIST   (b) CIFAR-100 Super-class

Figure 3: Bloxplot of prediction sets. We display the lengths of conformal sets that are not equal to 1 over all the test data.

not held in the prediction sets) of concept conformal sets with different miscoverage rates. From this table, we can find that our proposed method can guarantee a probability of error of at most ε and fully exploits its error margin to optimize efficiency. In contrast, the Naive baseline simply outputs all possible concepts using the raw softmax probabilities and does not provide theoretical guarantees. In addition, we show concept conformal sets when $\varepsilon = 0.1$ in Figure 2. Intuitively, a blurred image tends to have a large prediction set due to its uncertainty during testing, while a clear image tends to have a small set. Thus, a smaller set means higher confidence in the prediction, and a larger set indicates higher uncertainty regarding the predicted concepts.

Next, we investigate the effectiveness of the final prediction sets within the self-explaining model. We provide insights into the label coverage (the percentage of times the prediction set size is 1 and contains the true label) as well as the average set size. Table 2 reports the obtained experimental results when $\varepsilon = 0.1$. From this table, we can see that our proposed method consistently outperforms both the Naive and Vanilla baselines. Specifically, our proposed
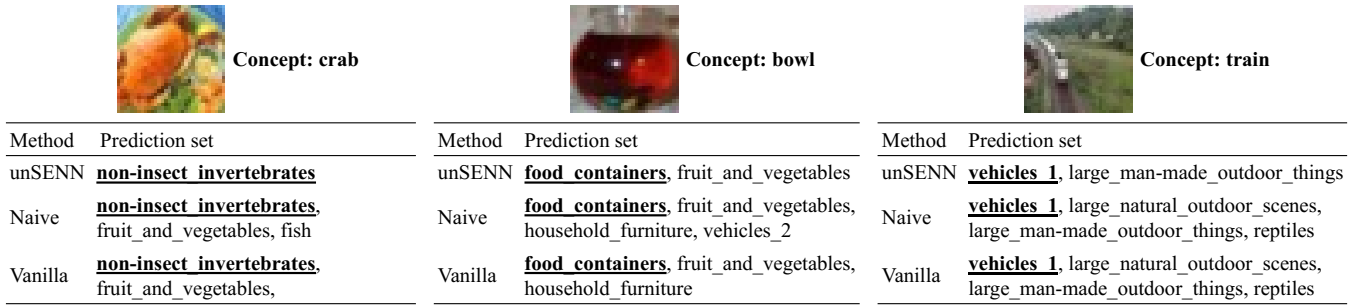
Method	Prediction set
unSENN	**non-insect_invertebrates**
Naive	**non-insect_invertebrates**, fruit_and_vegetables, fish
Vanilla	**non-insect_invertebrates**, fruit_and_vegetables,

Concept: crab

Method	Prediction set
unSENN	**food_containers**, fruit_and_vegetables
Naive	**food_containers**, fruit_and_vegetables, household_furniture, vehicles_2
Vanilla	**food_containers**, fruit_and_vegetables, household_furniture

Concept: bowl

Method	Prediction set
unSENN	**vehicles_1**, large_man-made_outdoor_things
Naive	**vehicles_1**, large_natural_outdoor_scenes, large_man-made_outdoor_things, reptiles
Vanilla	**vehicles_1**, large_natural_outdoor_scenes, large_man-made_outdoor_things, reptiles

Concept: train

Figure 4: Prediction conformal sets on CIFAR-100 Super-class. Phrases in bold and underlined mean true labels.
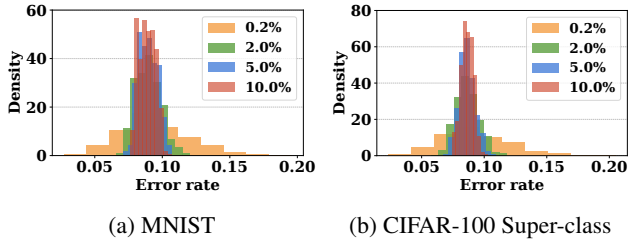

(a) MNIST   (b) CIFAR-100 Super-class

Figure 5: Distribution of error rate with different calibration sizes (splitting percentage of the original available data).

method achieves notably higher label coverage and much tighter confidence sets. A more comprehensive understanding of the conformal set characteristics is illustrated in Figure 3, where we display the lengths of conformal sets that are not equal to 1. We can find that the Vanilla baseline contains empty sets on MNIST, and the Naive baseline includes much looser prediction sets on the CIFAR-100 Super-class. In contrast, our proposed method generates reliable and tight prediction sets. Additionally, in Figure 4, we demonstrate prediction conformal sets when $\varepsilon = 0.1$. We can observe that all the methods correctly output the true label, while the prediction set of our proposed method is the tightest. From these reported experimental results, we can conclude that our proposed method effectively produces uncertainty prediction sets for the final predictions, leveraging the informative high-level basis concepts.

Further, we perform an ablation study concerning the calibration data size. In Figure 5, we illustrate the error rate distribution for the concept conformal sets with a miscoverage rate $\varepsilon = 0.1$ on MNIST and CIFAR-100 Super-class, using different calibration data sizes. It is evident that utilizing a larger calibration set can enhance the stability of the prediction set. Therefore, we incorporate 10% of the original available dataset for calibration without tuning, which also works well in our experiments.

Related Work

Currently, many self-explaining networks have been proposed (Koh et al. 2020; Havasi, Parbhoo, and Doshi-Velez 2022; Chauhan et al. 2023; Yuksekgonul, Wang, and Zou 2022; Alvarez Melis and Jaakkola 2018; Li et al. 2018; Kim

et al. 2023; Zarlenga et al. 2023). However, they fail to provide distribution-free uncertainty quantification for the two simultaneously generated prediction outcomes in self-explaining networks. On the other hand, to tackle the uncertainty issues, traditional methods (e.g., Bayesian approximation) for establishing confidence in prediction models usually make strict assumptions and have high computational cost properties. In this work, we build upon conformal prediction (Gibbs and Candes 2021; Chernozhukov, Wüthrich, and Yinchu 2018; Xu and Xie 2023; Tibshirani et al. 2019; Teng et al. 2022; Martinez et al. 2023; Ndiaye 2022; Lu et al. 2022; Jaramillo and Smirnov 2021; Liu et al. 2022), which do not require strict assumptions on the underlying data distribution and has been applied before to classification and regression problems to attain marginal coverage. However, our work is different from existing conformal works in that we first model explanation prediction uncertainties in both scenarios – with and without true explanations. We also transfer the explanation uncertainties to the final prediction uncertainties through designing new transfer functions and associated optimization frameworks for constructing the final prediction sets.

Conclusion

In this paper, we design a novel uncertainty quantification framework for self-explaining networks, which generates distribution-free uncertainty estimates for the two interconnected outcomes (i.e., the final predictions and their corresponding explanations) without requiring any additional model modifications. Specifically, in our approach, we first design the non-conformity measures to capture the relevant basis explanations within the interpretation layer of self-explaining networks, considering scenarios with and without true concepts. Then, based on the designed non-conformity measures, we generate explanation prediction sets whose size reflects the prediction uncertainty. To get better label prediction sets, we also develop novel transfer functions tailored to scenarios where self-explaining networks are trained with and without true concepts. By leveraging the adversarial attack paradigm, we further design novel optimization frameworks to construct final prediction sets from the designed transfer functions. We provide the theoretical analysis for the proposed method. Comprehensive experiments validate the effectiveness of our method.

Acknowledgments

References

Agarwal, C.; Krishna, S.; Saxena, E.; Pawelczyk, M.; Johnson, N.; Puri, I.; Zitnik, M.; and Lakkaraju, H. 2022. Openxai: Towards a transparent evaluation of model explanations. *Advances in Neural Information Processing Systems*, 35: 15784–15799.

Alvarez Melis, D.; and Jaakkola, T. 2018. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31.

Angelopoulos, A. N.; and Bates, S. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Bang, S.; Xie, P.; Lee, H.; Wu, W.; and Xing, E. 2021. Explaining a black-box by using a deep variational information bottleneck approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11396–11404.

Brunet, M.-E.; Anderson, A.; and Zemel, R. 2022. Implications of Model Indeterminacy for Explanations of Automated Decisions. *Advances in Neural Information Processing Systems*, 35: 7810–7823.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, 39–57. Ieee.

Chauhan, K.; Tiwari, R.; Freyberg, J.; Shenoy, P.; and Dvijotham, K. 2023. Interactive concept bottleneck models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 5948–5955.

Chernozhukov, V.; Wüthrich, K.; and Yinchu, Z. 2018. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Conference On learning theory*, 732–749. PMLR.

Cherubin, G.; Chatzikokolakis, K.; and Jaggi, M. 2021. Exact optimization of conformal predictors via incremental and decremental learning. In *International Conference on Machine Learning*, 1836–1845. PMLR.

Deng, L. 2012. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6): 141–142.

Elbaghdadi, O.; Hussain, A.; Hoenes, C.; and Bardarov, I. 2020. Self explaining neural networks: A review with extensions. *Fairness, Accountability, Confidentiality and Transparency in AI*.

Fischer, M.; Balunovic, M.; Drachsler-Cohen, D.; Gehr, T.; Zhang, C.; and Vechev, M. 2019. DL2: training and querying neural networks with logic. In *International Conference on Machine Learning*, 1931–1941. PMLR.

Gibbs, I.; and Candes, E. 2021. Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34: 1660–1672.

Han, T.; Srinivas, S.; and Lakkaraju, H. 2022. Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations. *Advances in Neural Information Processing Systems*, 35: 5256–5268.

Havasi, M.; Parbhoo, S.; and Doshi-Velez, F. 2022. Addressing leakage in concept bottleneck models. *Advances in Neural Information Processing Systems*, 35: 23386–23397.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hitzler, P.; and Sarker, M. 2022. Human-centered concept explanations for neural networks. *Neuro-Symbolic Artificial Intelligence: The State of the Art*, 342(337): 2.

Hong, J.; and Pavlic, T. P. 2023. Concept-Centric Transformers: Concept Transformers with Object-Centric Concept Learning for Interpretability. *arXiv preprint arXiv:2305.15775*.

Huai, M.; Liu, J.; Miao, C.; Yao, L.; and Zhang, A. 2022. Towards automating model explanations with certified robustness guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6935–6943.

Huai, M.; Wang, D.; Miao, C.; and Zhang, A. 2020. Towards interpretation of pairwise learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4166–4173.

Huang, W. R.; Geiping, J.; Fowl, L.; Taylor, G.; and Goldstein, T. 2020. Metapoison: Practical general-purpose clean-label data poisoning. *Advances in Neural Information Processing Systems*, 33: 12080–12091.

Jaramillo, W. L.; and Smirnov, E. 2021. Shapley-value based inductive conformal prediction. In *Conformal and Probabilistic Prediction and Applications*, 52–71. PMLR.

Kim, E.; Jung, D.; Park, S.; Kim, S.; and Yoon, S. 2023. Probabilistic Concept Bottleneck Models. *arXiv preprint arXiv:2306.01574*.

Koh, P. W.; Nguyen, T.; Tang, Y. S.; Mussmann, S.; Pierson, E.; Kim, B.; and Liang, P. 2020. Concept bottleneck models. In *International conference on machine learning*, 5338–5348. PMLR.

Li, J.; Kuang, K.; Li, L.; Chen, L.; Zhang, S.; Shao, J.; and Xiao, J. 2021. Instance-wise or class-wise? a tale of neighbor shapley for concept-based explanation. In *Proceedings of the 29th ACM International Conference on Multimedia*, 3664–3672.

Li, O.; Liu, H.; Chen, C.; and Rudin, C. 2018. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Li, W.; Feng, X.; An, H.; Ng, X. Y.; and Zhang, Y.-J. 2020. Mri reconstruction with interpretable pixel-wise operations using reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, 792–799.

Liu, M.; Ding, L.; Yu, D.; Liu, W.; Kong, L.; and Jiang, B. 2022. Conformalized Fairness via Quantile Regression. *Advances in Neural Information Processing Systems*, 35: 11561–11572.

Lu, C.; Lemay, A.; Chang, K.; Höbel, K.; and Kalpathy-Cramer, J. 2022. Fair conformal predictors for applications in medical imaging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 12008–12016.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Martinez, J. A.; Bhatt, U.; Weller, A.; and Cherubin, G. 2023. Approximating Full Conformal Prediction at Scale via Influence Functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6631–6639.

Nam, W.-J.; Gur, S.; Choi, J.; Wolf, L.; and Lee, S.-W. 2020. Relative attributing propagation: Interpreting the comparative contributions of individual units in deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, 2501–2508.

Ndiaye, E. 2022. Stable conformal prediction sets. In *International Conference on Machine Learning*, 16462–16479. PMLR.

Rajagopal, D.; Balachandran, V.; Hovy, E.; and Tsvetkov, Y. 2021. Selfexplain: A self-explaining architecture for neural text classifiers. *arXiv preprint arXiv:2103.12279*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Rigotti, M.; Miksovic, C.; Giurgiu, I.; Gschwind, T.; and Scotton, P. 2021. Attention-based interpretability with concept transformers. In *International Conference on Learning Representations*.

Rudin, C. 2018. Please stop explaining black box models for high stakes decisions. *Stat*, 1050: 26.

Sesia, M.; and Favaro, S. 2022. Conformal frequency estimation with sketched data. *Advances in Neural Information Processing Systems*, 35: 6589–6602.

Sinha, S.; Huai, M.; Sun, J.; and Zhang, A. 2023. Understanding and enhancing robustness of concept-based models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 15127–15135.

Slack, D.; Hilgard, A.; Singh, S.; and Lakkaraju, H. 2021. Reliable post hoc explanations: Modeling uncertainty in explainability. *Advances in neural information processing systems*, 34: 9391–9404.

Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; and Lakkaraju, H. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180–186.

Teng, J.; Tan, Z.; and Yuan, Y. 2021. T-sci: A two-stage conformal inference algorithm with guaranteed coverage for cox-mlp. In *International Conference on Machine Learning*, 10203–10213. PMLR.

Teng, J.; Wen, C.; Zhang, D.; Bengio, Y.; Gao, Y.; and Yuan, Y. 2022. Predictive inference with feature conformal prediction. *arXiv preprint arXiv:2210.00173*.

Tibshirani, R. J.; Foygel Barber, R.; Candes, E.; and Ramdas, A. 2019. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32.

Xu, C.; and Xie, Y. 2023. Sequential predictive conformal inference for time series. In *International Conference on Machine Learning*, 38707–38727. PMLR.

Yao, L.; Li, Y.; Li, S.; Liu, J.; Huai, M.; Zhang, A.; and Gao, J. 2022. Concept-Level Model Interpretation From the Causal Aspect. *IEEE Transactions on Knowledge and Data Engineering*.

Yeh, C.-K.; Kim, B.; Arik, S.; Li, C.-L.; Pfister, T.; and Ravikumar, P. 2020. On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems*, 33: 20554–20565.

Yuksekgonul, M.; Wang, M.; and Zou, J. 2022. Post-hoc Concept Bottleneck Models. In *The Eleventh International Conference on Learning Representations*.

Zarlenga, M. E.; Shams, Z.; Nelson, M. E.; Kim, B.; and Jamnik, M. 2023. TabCBM: Concept-based Interpretable Neural Networks for Tabular Data. *Transactions on Machine Learning Research*.

Zhang, Y.; Gao, N.; and Ma, C. 2023. Learning to Select Prototypical Parts for Interpretable Sequential Data Modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 6612–6620.

Zhao, C.; Qian, W.; Shi, Y.; Huai, M.; and Liu, N. 2023. Automated Natural Language Explanation of Deep Visual Neurons with Large Models. *arXiv preprint arXiv:2310.10708*.

Zhao, S.; Ma, X.; Wang, Y.; Bailey, J.; Li, B.; and Jiang, Y.-G. 2021. What do deep nets learn? class-wise patterns revealed in the input space. *arXiv preprint arXiv:2101.06898*.