Uncertainty Considerations of Explainable AI in Data-Driven Systems

Fatima Rabia Yapicioglu[1,2]

[1]Department of Computer Science and Engineering, University of Bologna, Bologna, Italy
[2]Marketing and Sales, Automobili Lamborghini S.p.A., Sant'Agata Bolognese, Italy

Abstract

Artificial Intelligence (AI) systems are increasingly relied upon in high-stakes domains such as healthcare, finance, and autonomous driving, as well as in high-value commercial applications like luxury automotive design and exclusive financial services, where decision-making must be both accurate and trustworthy. However, the opaque nature of many AI models raises concerns about transparency and accountability, driving the development of Explainable AI (XAI) techniques to foster trust. While these methods aim to improve interpretability, questions persist regarding the reliability and certainty of these explanations, particularly under varying conditions and sources of uncertainty. This underscores the need for robust trust measures to assess the validity and consistency of AI-generated explanations across different contexts. Consequently, the question shifts from *"Can I trust this model?"* to *"To what extent can I trust the explanations and the reasoning behind the model's decisions?"*—emphasizing the importance of reliable frameworks for explainability. To reliably quantify uncertainty in AI-generated predictions, we integrate conformal prediction, a distribution-free, model-agnostic framework that constructs prediction sets with statistically valid coverage guarantees, ensuring that the true outcome is included with a user-specified probability. By adapting to different tasks and data distributions, conformal prediction provides a robust foundation for uncertainty measurement and enables the generation of consistent, uncertainty-aware explanations across varying conditions. We term this approach "*uncertainty-aware explanations*", providing systematic methods to assess the trustworthiness of AI insights in diverse contexts, including time series forecasting, classification, and other data-driven tasks. By addressing the relationship between uncertainty and explainability, this work aims to enhance the reliability of AI-driven decision-making in high-stakes environments.

Keywords

Explainable Artificial Intelligence, Uncertainty-Awareness, Certainty in Explanations, Trustworthy AI

1. Context and Motivation

As Artificial Intelligence (AI) systems become increasingly complex, explaining their decisions becomes more challenging, leading to concerns about trust and reliability. This has prompted the development of Explainable AI (XAI) to enhance transparency by providing human-interpretable explanations for AI-driven decisions. XAI methods produce different types of explanations depending on the task, affecting their applicability across domains. For example, a practitioner analyzing ECG signals to assess a patient's risk of developing cardiac disease requires retrospective analysis to identify key time intervals contributing to the prediction (time series forecasting) and to rank other influential factors such as diet, weight, and physical activity (time series classification). Despite advancements in Explainable AI (XAI), there is a lack of robust methods to quantify the uncertainty in AI-generated explanations, leading to potential overreliance on explanations provided in critical domains. As a result, a critical question arises: *Can we trust these explanations, and if so, to what extent?* This highlights the need for rigorous evaluation frameworks to assess the reliability, consistency, and validity of AI-generated explanations across various contexts [1].

Uncertainty quantification (UQ) plays a critical role in AI systems, as it provides a measure of confidence in the model's predictions, helping practitioners assess the reliability of outputs, especially in high-stakes applications [2]. A promising approach to enhancing the reliability of AI-generated

Late-breaking work, Demos and Doctoral Consortium, colocated with The 3rd World Conference on eXplainable Artificial Intelligence: July 09–11, 2025, Istanbul, TR
fatima.yapicioglu2@unibo.it (F. R. Yapicioglu)
0000-0001-5888-445X (F. R. Yapicioglu)

explanations is conformal prediction, a statistical framework that offers valid uncertainty quantification with formal guarantees. In a classification task, conformal prediction provides a *prediction set*—a collection of possible labels for a new instance—accompanied by a confidence level. Instead of assigning a single label, the model offers a set of labels guaranteed to contain the true label with a specified probability, such as 90%. This enables practitioners to understand both the model's most likely prediction and the associated uncertainty, leading to more informed and reliable decision-making. By producing prediction sets that adapt to the uncertainty present in the data, conformal prediction offers well-calibrated measures of confidence and paves the way for reliable *uncertainty-aware explanations* [3].

Uncertainty-aware explanations in XAI enhance interpretability by revealing both the reasoning behind a model's outputs and the confidence in those outputs. In this work, they are defined as explanations that capture how variations in uncertainty influence the trustworthiness of predictions. For example, in financial risk assessment, such an explanation might show that a loan applicant's risk score is less reliable due to missing income data, with uncertainty rising by 30% compared to complete cases. This approach supports practitioners in evaluating the robustness of AI insights and making informed decisions under uncertainty.

The lack of uncertainty-aware explanations in AI can cause major issues, especially in high-stakes areas like healthcare and autonomous driving [4, 5]. For instance, without uncertainty quantification, a healthcare model may overestimate a patient's risk, leading to incorrect diagnoses due to incorrect reasoning, or an autonomous vehicle may make decisions without knowing the level of uncertainty, increasing accident risk. By integrating conformal prediction, we can generate explanations that not only identify key features but also quantify the confidence in these attributions, ensuring more reliable and transparent AI decision-making across applications.

2. Background

2.1. Model-Agnostic and Post-hoc Explainability

Post-hoc explainability refers to techniques applied after a machine learning model has been trained, aiming to interpret its predictions without altering the underlying model structure [6]. These methods are particularly valuable in high-stakes domains such as healthcare and finance, where understanding the rationale behind predictions is crucial for trust and accountability [7]. Post-hoc approaches include feature importance methods, surrogate models, and visualization techniques, which help uncover the decision-making process of complex models like deep neural networks or ensemble methods [8].

Model-agnostic explainability, a subset of post-hoc methods, is designed to be applicable to any machine learning model, regardless of its architecture or complexity [6]. There are two approaches: global explainability for overall patterns and local explainability for individual predictions. Techniques such as LIME (Local Interpretable Model-agnostic Explanations) [8] and SHAP (SHapley Additive exPlanations) [9] are prominent examples. LIME approximates the behavior of a model locally by creating interpretable surrogate models, while SHAP leverages game theory to attribute prediction outcomes to individual features. These methods provide flexibility and transparency, making them widely adopted in practice.

The growing demand for explainability stems from regulatory requirements, such as the European Union's General Data Protection Regulation (GDPR), which emphasizes the right to explanation [10]. Additionally, model-agnostic methods enable practitioners to maintain high predictive performance while ensuring interpretability, bridging the gap between accuracy and transparency [11].

2.2. Uncertainty Quantification and Conformal Prediction

Uncertainty quantification (UQ) is a fundamental aspect of machine learning that focuses on measuring and interpreting the uncertainty associated with model predictions. This is particularly critical in high-stakes applications such as healthcare, autonomous systems, and financial forecasting, where decisions based on overconfident predictions can lead to severe consequences [12]. UQ methods aim to provide

probabilistic estimates, confidence intervals, or prediction intervals to convey the reliability of model outputs. These techniques can be broadly categorized into Bayesian approaches, ensemble methods, and evidential deep learning [13, 14]. For example, Bayesian neural networks quantify uncertainty by modeling distributions over model parameters, while ensemble methods leverage multiple models to estimate predictive variance [15].

Conformal prediction is a model-agnostic, non-parametric framework for uncertainty quantification (UQ) that provides valid confidence intervals without strong distributional assumptions [16]. Relying on the weaker exchangeability assumption rather than i.i.d., it calibrates prediction sets or intervals using a hold-out validation set to guarantee user-specified coverage (e.g., 95%) [17]. It applies to any model, including black-box architectures, and has recently been extended to time-series forecasting and high-dimensional data [18].

2.2.1. Mathematical Formulation of Conformal Prediction

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ denote a dataset, where x_i represents the input features and y_i represents the corresponding true label or value. Conformal prediction works as follows:

1. **Nonconformity Measure**: A nonconformity measure $s(x, y)$ quantifies how unusual a pair (x, y) is with respect to the model's predictions. For example, in regression, $s(x, y)$ could be the absolute residual $|y - \hat{y}|$, where $\hat{y}$ is the model's prediction.
2. **Calibration Set**: A hold-out calibration set $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^{m}$ is used to compute nonconformity scores $s_i = s(x_i, y_i)$ for each point in the calibration set.
3. **Prediction Interval Construction**: For a new input x_{new}, the conformal prediction framework constructs a prediction interval $C(x_{\text{new}})$ such that:

$$C(x_{\text{new}}) = \{y : s(x_{\text{new}}, y) \leq q_\alpha\},$$

where q_α is the $(1 - \alpha)$-th quantile of the nonconformity scores $\{s_i\}_{i=1}^{m}$. This ensures that the interval $C(x_{\text{new}})$ covers the true label y_{new} with probability $1 - \alpha$ (α is user-specified error rate).

2.2.2. Key Metrics in Conformal Prediction

1. **Coverage**: Coverage measures the proportion of true labels that fall within the prediction intervals or sets. For a dataset $\mathcal{D}$, the empirical coverage is defined as:

$$\text{Coverage} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(y_i \in C(x_i)), \tag{1}$$

where $\mathbb{I}(\cdot)$ is the indicator function. A well-calibrated conformal prediction framework ensures that the empirical coverage is approximately $1 - \alpha$.
2. **Set Size**: Set size measures the size of the prediction sets or intervals. For classification tasks, it is the number of labels in the prediction set, and for regression tasks, it is the width of the prediction interval. Smaller set sizes indicate more precise predictions, while larger set sizes reflect higher uncertainty.

2.2.3. Integration and Advantages of Conformal Prediction

Integrating UQ and conformal prediction into machine learning pipelines improves decision-making by offering insights into prediction reliability, crucial in safety-critical applications where overconfidence can lead to catastrophic outcomes [19]. In healthcare, conformal prediction provides confidence intervals for patient outcomes, aiding clinicians in making informed decisions. In autonomous systems, UQ assesses prediction reliability in dynamic environments.

These methods align with the growing focus on transparency and robustness in AI, supported by regulatory frameworks and industry standards [20]. Conformal prediction offers valid coverage

guarantees without distribution assumptions, is model-agnostic, and is applicable to various tasks, including time-series forecasting and high-dimensional data, with recent extensions like split and adaptive conformal prediction [18].

3. Related Work

Related work linking explainability and conformal prediction remains limited. [21] propose CONFIDERAI, refining rule-based classifiers by combining conformal prediction with explainable ML for improved reliability. [22] introduces CONFINE, a framework for interpretable neural networks with robust uncertainty estimates. [23] explores oracle coaching to generate valid, efficient conformal classifiers optimized for specific test sets. [24] compares frequentist, Bayesian, and conformal uncertainty estimation, highlighting conformal methods for trustworthy confidence sets in model explanations. [25] apply XAI to cardiovascular risk prediction in COPD patients, comparing counterfactual methods and proposing counterfactual conformity for validation. [26] presents a conformal prediction-based framework for interpreting unsupervised node representations in graphs. Most relevant to this PhD is Calibrated Explanations (CE) [27], which provides stable, model-agnostic local feature importance maps with uncertainty quantification via Venn-Abers predictors [28]. In contrast, this work uses a perturbation-based, post-hoc, model-agnostic approach with classical conformal prediction, tailored to specific tasks and analyzing how predictive uncertainty shifts under varying calibration sets and systematic noise.

4. Research Questions and Objectives

Following research questions (RQ) have been proposed for this research:

1. How can uncertainty in AI-predictions be effectively quantified to enhance trust and reliability in decision-making?
2. Which evaluation measures are needed to assess the validity/performance of conformal prediction, and how can we leverage them to produce uncertainty-aware explanations?
3. How do uncertainty-aware explanations generated by our proposed frameworks enhance decision-making and compare to conventional explainers across diverse real-world scenarios?

To address **RQ1**, we begin by quantifying uncertainty in AI predictions using the flexible framework of *conformal prediction*. In classification tasks, this involves *prediction sets* with varying *confidence levels*, while in regression, it involves *prediction intervals* around outputs. Additionally, scalar uncertainty measures—such as variance from ensembles, dropout, input perturbations, or adversarial modifications—provide adaptable, task-specific metrics [29]. This phase identifies the most suitable uncertainty quantification methods across different tasks.

For **RQ2**, we extend conformal prediction to evaluate and communicate uncertainty in AI-generated explanations across tasks like *classification*, *time series forecasting*, and *clustering*. A key property of conformal prediction is that its coverage, while guaranteed on average (Equation 1), varies with calibration sets. We will explore how perturbing training or calibration data affects uncertainty and model performance, aiming to integrate these effects into *reliable* and *transparent* explanation frameworks.

Uncertainty metrics will be task-specific: in classification, we analyze *prediction set size* and *coverage* (Section 2.2.2); in forecasting, we assess changes in *confidence interval bounds* (Equation 3). Perturbing input features helps recalibrate models and track shifts in uncertainty. We also explore both *local* and *global* explainability.

In **RQ3**, we compare our uncertainty-aware explanation frameworks with SHAP, LIME, Saliency Maps, and Integrated Gradients. *Effectiveness* is tested via ablation of the top-ranked feature or segment; *robustness* by varying conformal prediction confidence levels; and *faithfulness* by comparison with inherently explainable models. These evaluations integrate uncertainty to enhance transparency and reliability.

5. Research Approach, Methods, and Rationale for Testing the Hypothesis

This research aims to develop task-specific uncertainty-aware explanations within a conformal prediction framework. Using a modular approach for classification, forecasting, and other tasks, it systematically measures and explains feature or segment contributions to predictive uncertainty.

5.1. Approach and Methods

For each task (e.g., classification, regression, time-series forecasting), the conformal prediction framework is tailored to produce prediction sets or intervals that quantify uncertainty, ensuring task-specific validity and interpretability [17].

Classification Tasks Given an input x and error rate $\alpha \in (0, 1)$, the prediction set is:

$$C(x) = \{y \in \mathcal{Y} \mid s(x, y) \geq \tau_\alpha\}, \tag{2}$$

where $\mathcal{Y}$ is the label set, $s(x, y)$ is a conformity score, and τ_α is a threshold ensuring coverage of $1 - \alpha$.

Regression Tasks For a predicted value $\hat{y}(x)$, the prediction interval is:

$$I(x) = [\hat{y}(x) - \tau_\alpha, \hat{y}(x) + \tau_\alpha], \tag{3}$$

ensuring the interval captures the true value with probability $1 - \alpha$.

We assess feature or segment contributions to uncertainty by applying systematic perturbations and measuring changes in *Coverage* and *Set-size* (Section 2.2.2). For *classification*, individual features are perturbed and models recalibrated; for *forecasting*, PELT segmentation [30] is used, and perturbing each segment reveals its impact on interval bounds (Equation 3). Comparisons with the unmodified baseline identify the main sources of predictive uncertainty.

5.2. Rationale and Testing

The rationale for using conformal prediction for producing uncertainty-aware explanations is its ability to provide reliable uncertainty estimates independent of data distribution. However, despite its robust uncertainty quantification, conformal prediction often lacks systematic methods for explaining the sources of uncertainty and adapting to various tasks.

We hypothesize that perturbing features or segments and examining their effect on uncertainty metrics can yield meaningful, interpretable uncertainty-aware explanations. This is validated on diverse datasets to ensure robustness and generalizability. Effectiveness is assessed through ablation studies, where significant features or intervals are removed and performance is remeasured. We further evaluate *robustness* by varying confidence levels $(1 - \alpha)$, as defined in Section 5.1, and *faithfulness* by comparing our results to explanations from intrinsically interpretable models, verifying that they reflect genuine uncertainty sources rather than artifacts of the method.

6. Results and Contributions to Date

In our research, we developed the global explainability framework *ConformaSight* [4], based on conformal prediction, to address prediction set-type outputs across classifiers. The framework generates a global feature importance table, making it easy for non-experts to identify factors affecting conformal metrics like coverage and set-size. In effectiveness tests, selecting the top 7 features by each explainer and retraining models resulted in a 0.7% improvement over SHAP and Permutation [31].

We contributed to Fast Calibrated Explanations (FCE) [5], which combines ConformaSight's perturbation techniques with Calibrated Explanations (CE) [27] to deliver rapid, uncertainty-aware explanations.

Applicable to classification and thresholded regression, FCE provides probabilistic outputs while preserving uncertainty quantification, achieving up to $19\times$ faster regression, over $75\times$ faster than calibrated LIME, and $200\times$ faster than calibrated SHAP. Having developed uncertainty-aware explanations for classification and regression tasks, we are now extending our research to time-series forecasting. We present *ConformaSegment*, a segment-based explanation framework that identifies, segments, and weights time-series intervals by their decision importance. In ablation studies on the most influential segment, *ConformaSegment* outperformed *Saliency Maps* [32] with a 42% average R^2 gain and 25.73% higher prediction interval coverage, and *Integrated Gradients* [33] with an 18% R^2 gain and 40.15% higher coverage.

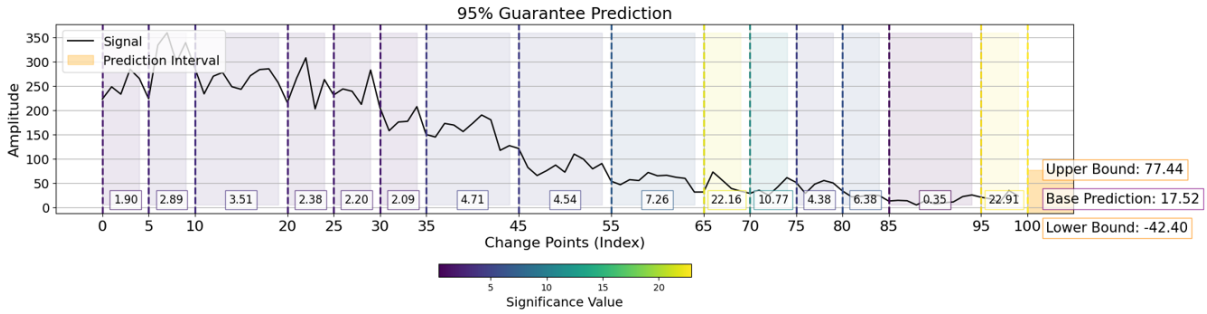

Figure 1: ConformaSegment: Segment-based Feature Importance by using Electric Power Consumption Dataset. Vertical dash-lines are the change points. Below rectangles between change points are the importance weights [34].

In summary, the contributions up to date are as follows:

1. **Leveraged conformal prediction to generate uncertainty-aware explanations for tabular data classification (ConformaSight):** We designed a framework to identify which features contribute most to predictive uncertainty when subjected to significant perturbations, potentially causing the model to make incorrect predictions. The framework shows how calibration set perturbations influence prediction set outcomes, highlighting their impact on model performance.
2. **Contributed FCE for rapid uncertainty-aware explanations for tabular data classification and regression:** We proposed a method designed for generating faster, uncertainty-aware explanations by incorporating perturbation techniques from ConformaSight into the core elements of CE. This method boosts computational efficiency for real-time use while preserving uncertainty quantification in classification and probabilistic regression.
3. **Extended conformal prediction to generate uncertainty-aware explanations for time-series forecasting (ConformaSegment):** We adapted our framework to time-series forecasting tasks, focusing on identifying the most critical time segments that contribute to predictive uncertainty, thereby influencing the accuracy of the forecasted values.

7. Expected Next Steps and Final Contribution to Knowledge

This PhD research aims to enhance trust in model decision-making by identifying key factors driving significant changes in model uncertainty. We explore how conformal prediction, which provides statistically guaranteed prediction sets with user-specified coverage, can be leveraged to generate uncertainty-aware explanations. Our goal is to develop a family of post-hoc, model-agnostic frameworks designed to produce reliable and interpretable explanations while advancing the transparency of conformal prediction-based explainability methods. Next, we aim to extend these frameworks to anomaly detection, synthetic data generation, and clustering, advancing transparent and generalizable uncertainty-aware explainability.

8. Acknowledgments

Fatima Rabia is a PhD student at DISI, University of Bologna, funded by PNRR (n. 9990) and Automobili Lamborghini S.p.A., Italy.

9. Declaration on Generative AI

The author has used ChatGPT-4o exclusively for grammar checking and rephrasing; the author originally produced all content.

References

[1] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence, Information fusion 99 (2023) 101805.

[2] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, S. Nahavandi, A review of uncertainty quantification in deep learning: Techniques, applications and challenges, Information Fusion 76 (2021) 243–297. URL: https://www.sciencedirect.com/science/article/pii/S1566253521001081. doi:https://doi.org/10.1016/j.inffus.2021.05.008.

[3] Y. Zhang, Q. V. Liao, R. K. Bellamy, Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making, in: Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020, pp. 295–305.

[4] F. R. Yapicioglu, A. Stramiglio, F. Vitali, Conformasight: Conformal prediction-based global and model-agnostic explainability framework, in: World Conference on Explainable Artificial Intelligence, Springer, 2024, pp. 270–293.

[5] T. Löfström, F. R. Yapicioglu, A. Stramiglio, H. Löfström, F. Vitali, Fast calibrated explanations: Efficient and uncertainty-aware explanations for machine learning models, arXiv preprint arXiv:2410.21129 (2024).

[6] C. Molnar, G. Casalicchio, B. Bischl, Interpretable machine learning–a brief history, state-of-the-art and challenges, in: Joint European conference on machine learning and knowledge discovery in databases, Springer, 2020, pp. 417–431.

[7] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (2018) 1–42.

[8] M. T. Ribeiro, S. Singh, C. Guestrin, " 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, 2016, pp. 1135–1144.

[9] S. Lundberg, A unified approach to interpreting model predictions, arXiv preprint arXiv:1705.07874 (2017).

[10] B. Goodman, S. Flaxman, European union regulations on algorithmic decision-making and a "right to explanation", AI magazine 38 (2017) 50–57.

[11] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information fusion 58 (2020) 82–115.

[12] R. C. Smith, Uncertainty Quantification: Theory, Implementation, and Applications, SIAM, 2013.

[13] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision?, Advances in Neural Information Processing Systems 30 (2017).

[14] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, Advances in Neural Information Processing Systems 30 (2017).

[15] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in

deep learning, in: Proceedings of the 33rd International Conference on Machine Learning, PMLR, 2016, pp. 1050–1059.
[16] V. Vovk, A. Gammerman, G. Shafer, Algorithmic Learning in a Random World, Springer, 2005.
[17] G. Shafer, V. Vovk, A tutorial on conformal prediction, Journal of Machine Learning Research 9 (2008) 371–421.
[18] A. N. Angelopoulos, S. Bates, Conformal prediction: A gentle introduction, Foundations and Trends in Machine Learning 16 (2023) 1–100.
[19] Z. Ghahramani, Probabilistic machine learning and artificial intelligence, Nature 521 (2015) 452–459.
[20] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning: Limitations and Opportunities, fairmlbook. org, 2020.
[21] S. Narteni, A. Carlevaro, F. Dabbene, M. Muselli, M. Mongelli, Confiderai: a novel conformal interpretable-by-design score function for explainable and reliable artificial intelligence, arXiv preprint arXiv:2309.01778 (2023).
[22] L. Huang, S. Lala, N. K. Jha, Confine: Conformal prediction for interpretable neural networks, arXiv preprint arXiv:2406.00539 (2024).
[23] U. Johansson, T. Löfström, H. Boström, C. Sönströd, Interpretable and specialized conformal predictors, in: A. Gammerman, V. Vovk, Z. Luo, E. Smirnov (Eds.), Proceedings of the Eighth Symposium on Conformal and Probabilistic Prediction and Applications, volume 105 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 3–22. URL: https://proceedings.mlr.press/v105/johansson19a.html.
[24] C. Marx, Y. Park, H. Hasson, Y. Wang, S. Ermon, L. Huan, But are you sure? an uncertainty-aware perspective on explainable ai, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2023, pp. 7375–7391.
[25] M. Lenatti, A. Carlevaro, A. Guergachi, K. Keshavjee, M. Mongelli, A. Paglialonga, Estimation and conformity evaluation of multi-class counterfactual explanations for chronic disease prevention, IEEE Journal of Biomedical and Health Informatics (2024).
[26] H. Park, Providing post-hoc explanation for node representation learning models through inductive conformal predictions, IEEE Access 11 (2022) 1202–1212.
[27] H. Löfström, T. Löfström, U. Johansson, C. Sönströd, Calibrated explanations: With uncertainty information and counterfactuals, Expert Systems with Applications 246 (2024) 123154.
[28] V. Vovk, I. Petej, V. Fedorova, Large-scale probabilistic predictors with and without guarantees of validity, Advances in Neural Information Processing Systems 28 (2015).
[29] A. N. Angelopoulos, S. Bates, A gentle introduction to conformal prediction and distribution-free uncertainty quantification, arXiv preprint arXiv:2107.07511 (2021).
[30] R. Killick, P. Fearnhead, I. A. Eckley, Optimal detection of changepoints with a linear computational cost, Journal of the American Statistical Association 107 (2012) 1590–1598. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.2012.737745. doi:10.1080/01621459.2012.737745.
[31] A. Altmann, L. Tolosana-Delgado, O. E. Jensen, C. Wiuf, Permutation importance: a corrected feature importance measure, Bioinformatics 26 (2010) 1340–1347. URL: https://academic.oup.com/bioinformatics/article/26/10/1340/193348. doi:10.1093/bioinformatics/btq134.
[32] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the 34th International Conference on Machine Learning (ICML), 2017, pp. 3319–3328. URL: https://arxiv.org/abs/1703.01365.
[33] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, in: Proceedings of the 2nd International Conference on Learning Representations (ICLR), 2014. URL: https://arxiv.org/abs/1312.6034.
[34] D. Dua, C. Graff, UCI Machine Learning Repository: Individual Household Electric Power Consumption Dataset, 2019. URL: https://archive.ics.uci.edu/ml/datasets/individual+household+electric+power+consumption.