Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 06, June - 2024

Evaluating Conformal Prediction Techniques in Adversarial Environments: A Comparative Study

Bhargava Kumar
Independent Researcher
Columbia Univ Alum

Tejaswini Kumar
Independent Researcher
Columbia Univ Alum

Hitesh Patel
Independent Researcher
NYU Alum

***Abstract*— Conformal prediction is a framework in statistics that provides precise measures of uncertainty for machine learning predictions. This framework ensures that prediction sets or intervals have a specified probability of containing the true outcome. This property is especially important in adversarial settings, where models are susceptible to malicious attacks that can result in incorrect and overconfident predictions. In this paper, we conduct a thorough comparative analysis of the most advanced conformal prediction methods designed to improve robustness against adversarial attacks.**

Our evaluation encompasses five noteworthy techniques: "The Pitfalls and Promise of Conformal Inference Under Adversarial Attacks," "Robust Yet Efficient Conformal Prediction Sets," "Verifiably Robust Conformal Prediction (VRCP)," "Adversarially Robust Conformal Prediction," and "Calibrated Multi-probabilistic Prediction as a Defense Against Adversarial Attacks." Our examination delves into critical aspects, including computational efficiency, resilience to diverse types of adversarial attacks, practical implementation challenges, and theoretical guarantees of coverage in adversarial settings.

This study presents a comprehensive analysis of conformal prediction methods through empirical evaluations on benchmark datasets and theoretical examinations. The results of this research provide valuable insights into the practical implications of using these methods in real-world adversarial scenarios. Our findings shed light on the trade-offs between robustness and efficiency of each method and highlight the strengths and limitations of each approach. This paper contributes to the understanding of conformal prediction's role in enhancing model robustness and offers guidance for future research in this area.

***Keywords*— Conformal prediction, Adversarial attacks, Robustness, Uncertainty quantification**

I. INTRODUCTION

Conformal prediction serves as a robust statistical methodology that delivers prediction sets or intervals with a guaranteed coverage probability. In contrast to conventional prediction models, which furnish point estimates, conformal prediction ensures that the true outcome is contained within the anticipated set with a predetermined confidence level. The unique feature of conformal prediction makes it indispensable in applications that demand reliable uncertainty quantification, such as medical diagnoses, financial forecasting, and autonomous systems. Conformal prediction capitalizes on past data to calibrate the uncertainty in predictions, thereby adapting to the underlying data distribution and model, presenting a versatile and model-agnostic approach to uncertainty estimation.

The importance of building robust machine learning models has become increasingly apparent, particularly in light of adversarial attacks. Adversarial attacks involve intentionally manipulated input data, which can cause models to produce incorrect or highly confident predictions. In critical applications such as autonomous driving, adversarial examples can lead to catastrophic failures, while in healthcare, they can result in serious misdiagnoses. Therefore, it is crucial to ensure that models are robust against such attacks, as this guarantees that they remain reliable and trustworthy, even when subjected to malicious perturbations. Integrating conformal prediction methods with adversarial robustness techniques not only provides predictions but also offers insights into the model's confidence under adversarial conditions, thereby enhancing the overall reliability of the system.

The present paper has as its objective a comprehensive comparative analysis of various conformal prediction methods designed to enhance robustness against adversarial attacks. Our focus is on four prominent methods: Adversarially Robust Conformal Prediction [3], The Pitfalls and Promise of Conformal Inference Under Adversarial Attacks [4], Verifiably Robust Conformal Prediction (VRCP) [5], Robust Yet Efficient Conformal Prediction Sets [6]. The aim of this study is to evaluate these methods in terms of their computational efficiency, robustness to different types of adversarial attacks, practical implementation challenges, and theoretical coverage guarantees. The objective is to highlight the strengths and limitations of each method, with the aim of providing insights into their practical applicability and guiding future research directions in the development of robust conformal prediction models. This comparative study emphasizes the importance of robust conformal prediction methods in adversarial settings and offers a pathway for enhancing the resilience and reliability of machine learning models in critical applications.

II. LITERATURE REVIEW

Conformal prediction (CP) has garnered considerable attention in recent times due to its capacity for furnishing dependable uncertainty estimates. The foundational work on conformal prediction may be traced back to [1], who expounded upon the concept and introduced methods for delivering valid prediction intervals for any specified significance level. This early work laid the groundwork for the subsequent development of conformal prediction techniques across diverse domains, thereby underscoring the model-agnostic nature of CP.

Given the growing utilization of machine learning models in critical applications, the robustness of these models against

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 06, June - 2024

adversarial attacks has emerged as a vital area of study. Adversarial attacks, which involve minute modifications to input data that can result in significant errors in model predictions, have underscored the necessity for models that not only exhibit strong performance but also maintain dependability under adversarial conditions. Pioneering research conducted by[2] revealed the susceptibility of deep neural networks to adversarial attacks, thereby catalyzing the investigation of defensive measures.

With the increasing importance of conformal prediction, adversarial robustness has become a central concern. [3] addresses this issue by combining conformal prediction with randomized smoothing, thereby ensuring coverage guarantees even in the presence of adversarial perturbations. The outcomes of this technique indicate that incorporating Gaussian noise into the non-conformity scores can effectively account for adversarial interference.

The importance of incorporating adversarial training in order to maintain the reliability of conformal predictions was emphasized by one of the seminal works in this field [4]. This study demonstrated that standard conformal prediction methods are prone to failure under adversarial conditions unless they are adversarially trained.

The relevance of neural network verification in ensuring coverage guarantees under adversarial attacks was demonstrated by subsequent research. This approach offered two variants: VRCP through Robust Inference and VRCP through Robust Calibration, each presenting distinct strategies for achieving robust prediction sets. These methods represent a significant advance in the literature on robust conformal prediction [5].

Further exploration into the computational challenges associated with robust conformal prediction methods revealed important findings. The authors introduced CDF-aware smoothing techniques that improve robustness while maintaining computational efficiency, thereby making it possible to apply these methods in practical settings [6].

The advancements underscore the adaptive nature of conformal prediction techniques in light of the escalating challenges posed by adversarial attacks. Each iteration has constructively built upon previous efforts enhancing the effectiveness, efficiency, and versatility of conformal prediction in adversarial circumstances.

III. PROBLEM STATEMENT

The article at hand seeks to address the obstacle of guaranteeing the dependability of conformal prediction methods in adversarial circumstances. Conformal prediction offers precise uncertainty estimates by constructing prediction intervals that comprise the genuine outcome with a specified probability. However, these methods may become unreliable when the underlying machine learning models are subjected to adversarial attacks—deliberate perturbations designed to elicit incorrect predictions. The objective of the paper is to assess and contrast different conformal prediction techniques to determine their effectiveness in maintaining robustness under such adversarial conditions. The objective is to identify the trade-offs involved in balancing robustness and computational efficiency, comprehend the applicability of these methods in various contexts, and provide insights for future research to improve the generalization and practical implementation of conformal prediction techniques in adversarial settings.

IV. METHODOLIGIES

A. Adversarially Robust Conformal Prediction

1) Method Overview: The research proposes a technique that combines conformal prediction with randomized smoothing to increase resilience against adversarial attacks. This method aims to deliver valid prediction sets that maintain guaranteed coverage rates even when faced with adversarial perturbations. By employing randomized smoothing, the approach smoothes the nonconformity scores with Gaussian noise, thereby accounting for adversarial effects and offering robust prediction intervals.

2) Key Contributions:

a) Randomized Smoothing Integration: The key feature of this method is the utilization of randomized smoothing in conformal prediction. Randomized smoothing involves incorporating Gaussian noise into the inputs to create robust classifiers. By smoothing the non-conformity scores, the technique effectively constrains the Lipschitz constant, which helps in managing adversarial perturbations.

b) Finite-Sample Coverage Guarantees: The method offers theoretical guarantees that the prediction sets will encompass the true outputs with a high probability, even under adverse conditions. This is realized by deriving coverage bounds that hold for any data distribution with l_2-norm limited adversarial noise.

c) Empirical Validation on Benchmark Datasets: The study demonstrates the method's effectiveness on multiple benchmark datasets, such as CIFAR-10, CIFAR-100, and ImageNet. The results indicate that the proposed method provides robust prediction intervals across a range of adversarial attack scenarios, surpassing standard conformal prediction techniques in terms of robustness.

3) Limitations:

a) Computational Complexity: Adding randomized smoothing raises the computational burden of the method. The process of incorporating Gaussian noise and computing non-conformity scores with a smooth function is computationally intensive, particularly for large datasets and intricate models.

b) Assumption of l_2-Norm Bounded Noise: The robustness assurances are primarily established for l_2-norm bounded adversarial perturbations. Although this is a typical type of adversarial perturbation, the technique's performance against other types of adversarial attacks (e.g., l_∞-norm) is not comprehensively explored.

c) Dependency on Noise Parameters: The efficacy of randomized smoothing is reliant on the noise parameter selection. Choosing suitable noise levels is vital for striking a balance between robustness and accuracy; however, this process can be intricate and may necessitate extensive adjustments.

d) Scalability Concerns: Although the method exhibits robust performance on benchmark datasets, its scalability to extremely

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 06, June - 2024

large-scale real-world applications merits further investigation. The computational requirements of randomized smoothing could present challenges for its implementation in environments with limited computational resources.

B. The Pitfalls and Promise of Conformal Inference Under Adversarial Attacks

1) Method Overview: The research investigates the challenges and potential of conformal prediction (CP) under adversarial settings in machine learning. It highlights that standard CP methods fail to produce reliable prediction sets when models are not adversarially trained, particularly under l_∞-norm bounded attacks. Adversarial training (AT) is identified as crucial for CP effectiveness, although advanced AT variants like TRADES and MART often lead to larger prediction set sizes (PSS), indicating inefficiencies in uncertainty quantification despite improved robustness. To address this issue, authors come up with a novel approach.

2) Key Contributions:

a) Analysis of Standard CP Methods: The research examines the performance of classic CP methods in situations where they are subjected to adversarial attacks. The findings reveal that these methods are unable to generate informative prediction sets without undergoing adversarial training.

b) Importance of Adversarial Training: The scientists emphasize the importance of adversarial training to sustain the efficacy of CP techniques. They illustrate that models that are not trained adversarially produce excessively cautious prediction sets that lack practical value.

c) Uncertainty-Reducing Adversarial Training (AT-UR): In an effort to overcome the limitations of current techniques, the authors suggest a novel uncertainty-reducing adversarial training methodology. This approach integrates a Beta-weighting loss function and an entropy minimization regularizer to enhance the size of the prediction sets while guaranteeing stability.

d) Empirical Validation: The proposed method has been empirically validated on several image classification datasets, demonstrating its efficacy in producing dependable prediction sets under adversarial conditions as compared to conventional adversarial training techniques.

3) Limitations:

a) Dependence on Adversarial Training: The research underscores the inefficacy of CP methods without adversarial training. The necessity of adversarial training engenders complexity and exacerbates computational burdens, hence rendering the method impractical in situations wherein adversarial training proves unfeasible.

b) Prediction Set Size: Despite the introduction of the AT-UR method, the size of the prediction sets may still be larger than optimal, which could potentially affect the efficiency of the method. When the prediction sets are excessively large, the usefulness of the predictions may be diminished in practical settings where precise intervals are essential

c) Limited Attack Types: The research primarily focuses on l_∞-norm bounded attacks. Other types of adversarial attacks, such as those based on different norms or more sophisticated attack strategies, are not extensively covered.

d) Computational Overhead: The incorporation of entropy minimization and Beta-weighted loss in adversarial training amplifies the computational complexity. This could potentially impede the scalability of the methodology when applied to larger datasets or more intricate models.

C. Verifiably Robust Conformal Prediction (VRCP)

1) Method Overview: The present research unveils a novel approach, termed VRCP, which focuses on delivering precise prediction sets that maintain optimal coverage guarantees amidst adversarial circumstances. This technique employs neural network verification techniques to guarantee the reliability of conformal prediction intervals. VRCP offers two primary variants, namely VRCP-I and VRCP-C, which are designed to cater to various stages of the prediction process, thereby ensuring resilience against adversarial assaults.

2) Key Contributions:

a) Neural Network Verification Integration: The main achievement of VRCP lies in incorporating neural network verification methods within the conformal prediction framework, thereby enabling the development of verifiably robust prediction sets that remain resilient to adversarial perturbations. VRCP introduces two distinct methods to achieve robustness:

VRCP via Robust Inference (VRCP-I): The following approach entails employing neural network verification during the inference stage to determine conservative regions for the scores of test inputs. This method, known as VRCP-I, guarantees the dependability of the prediction sets in the face of adversarial circumstances.

VRCP via Robust Calibration (VRCP-C): This methodology employs neural network validation during the calibration phase to determine upper limits for calibration scores. By implementing a more stringent calibration criterion, VRCP-C guarantees that prediction sets preserve their coverage assurances without necessitating verification at the time of inference

b) Theoretical Guarantees: VRCP presents theoretical guarantees for the robustness of prediction sets. This method ensures that the prediction sets include the true test output with a high probability, even under the influence of adversarial attacks on the test inputs.

c) Empirical Validation: The method has been extensively tested on multiple datasets (including CIFAR10, CIFAR100, and TinyImageNet), demonstrating its ability to provide robust predictions even under various types of adversarial attacks. The results of these experiments indicate that VRCP successfully balances the need for robustness and the size of the prediction set, making it a practical choice for real-world scenarios.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 06, June - 2024

3) Limitations:

a) Computational Complexity: The main limitation of VRCP is the substantial computational overhead associated with verifying neural networks. Both VRCP-I and VRCP-C demand considerable computational resources, especially for large and intricate models, which can restrict the method's scalability and practicality in resource-constrained settings.

b) Norm-Bounded Perturbations: The primary objective of this method is to tackle norm-bounded adversarial perturbations, which have proven effective against common attack types. However, its robustness against more sophisticated forms of adversarial attacks has yet to be fully evaluated.

c) Dependency on Verification Accuracy: The dependability of VRCP is substantially influenced by the precision and comprehensiveness of the neural network validation process. Inefficient verification can result in suboptimal prediction sets, which in turn can compromise the overall resilience and assurance levels.

d) Implementation Complexity: Implementing neural network verification within the conformal prediction framework demands considerable knowledge and exertion. The intricacy of this process may present an obstacle to its adoption, especially in circumstances where practitioners lack extensive familiarity with neural network verification methodologies.

D. Robust Yet Efficient Conformal Prediction Sets

1) Method Overview: The present paper introduces a fresh strategy aimed at improving the effectiveness and practicality of conformal prediction techniques. The primary objective of this research is to tackle the computational issues related to devising sturdy conformal prediction sets that are suitable for practical applications. To achieve this aim, the authors propose CDF-conscious (Cumulative Distribution Function-conscious) smoothing procedures that ensure a balance between robustness and computational feasibility.

2) Key Contributions:

a) CDF-Aware Smoothing Techniques: The main focus of this research is the presentation of CDF-aware smoothing methods. These approaches are intended to adjust prediction intervals in a manner that is responsive to the distribution of non-conformity scores, thereby enhancing robustness without significantly raising computational demands.

b) Efficiency in Prediction Set Construction: Utilizing CDF-conscious smoothing techniques, the approach guarantees that the formation of prediction sets remains cost-effective. This is particularly crucial for situations requiring swift predictions, such as real-time decision-making mechanisms.

c) Empirical Validation: The method has been validated using several benchmark datasets, which demonstrate its ability to produce reliable prediction sets efficiently. The empirical results indicate that the proposed approach performs well under a variety of adversarial attack scenarios, maintaining an appropriate balance between robustness and practicality.

d) Theoretical Guarantees: The authors provide theoretical guarantees for the coverage and efficiency of their method. These guarantees ensure that the prediction sets remain reliable and practical even when facing adversarial perturbations.

3) Limitations:

a) Dependence on Data Distribution: The efficacy of CDF-conscious smoothing techniques is largely contingent upon the precise estimation of the underlying data distribution. In situations where the data distribution is either poorly comprehended or exhibits significant intricacy, the technique may not achieve optimal performance

b) Computational Cost of CDF-Aware Randomized Smoothing: While randomized smoothing is flexible and powerful, estimating empirical statistics requires a large number of Monte-Carlo samples, making it computationally expensive.

c) Limited Types of Adversarial Attacks: Similar to the evaluations conducted in other studies within this domain, the present method's robustness is primarily assessed against a restricted range of adversarial attack types (evasion, feature poisoning and label poisoning attack). However, future research should explore the performance of this method against more advanced or unforeseen attack strategies.

d) Implementation Complexity: Incorporating CDF-aware smoothing into current systems could prove to be a challenging task that necessitates substantial modifications and a comprehensive grasp of the technique. This may hinder its implementation in real-world scenarios, particularly in situations where financial, human, or technical resources are limited.

e) Fairness Issues: Even with group-conformal variants to equalize coverage across groups, unfairness can still manifest in the form of differences in set size among different groups, indicating a need for further study on the intersection of robustness and fairness.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 06, June - 2024

V. COMPARATIVE ANALYSIS OF CONFORMAL PREDICTION METHODS

Table 1: Comparative Analysis of Conformal Prediction Method

Criteria	Adversarially Robust Conformal Prediction	The Pitfalls and Promise of Conformal Inference Under Adversarial Attacks	Verifiably Robust Conformal Prediction (VRCP)	Robust Yet Efficient Conformal Prediction Sets
Efficiency				
Computational Complexity	Moderate, due to randomized smoothing	High due to adversarial training and additional loss functions	High, due to neural network verification	Moderate, with optimized CDF-aware smoothing
Scalability	Moderate scalability issues	Limited by computational demands	Limited scalability due to intensive verification	Better scalability, but challenged by high-dimensional data
Robustness				
Performance against Different Types of Adversarial Attacks	Robust against l_2-norm attacks	Effective mainly against l_∞-norm attacks	Robust against norm-bounded perturbations	Robust against various attack types
Empirical Results on Benchmark Datasets	Validated on CIFAR-10, CIFAR-100, ImageNet	Validated on multiple image classification datasets	Validated on CIFAR10, CIFAR100, and TinyImageNet for classification and other for regression	Tested on CIFAR-10, CIFAR-100, Cora-ML
Practicality				
Ease of Implementation	Moderate complexity with randomized smoothing	Complex due to adversarial training	Highly complex due to verification requirements	Moderate complexity with CDF-aware smoothing
Real-World Applicability	Applicable but challenging in resource-constrained environments	Suitable for environments with feasible adversarial training	Less practical due to high computational demands	Practical for real-world applications
Theoretical Guarantees				
Coverage Guarantees Under Adversarial Conditions	Finite-sample coverage guarantees under l_2-norm attacks	Provided under adversarial training scenarios	Strong guarantees with neural network verification	Theoretical guarantees based on CDF-aware smoothing
Limitations of Theoretical Bounds	Limited to l_2-norm attacks, may not extend to other perturbations	Dependent on the effectiveness of adversarial training	Dependent on precision of verification process	Influenced by the accuracy of CDF-aware smoothing

VI. DISCUSSION

The evaluation of conformal prediction techniques in adversarial settings unveils distinct advantages and constraints, providing a multifaceted understanding of their suitability in different situations. Each approach offers a unique balance between robustness and computational efficiency, making them suitable for a range of applications depending on specific requirements and limitations.

Choosing the right method depends on the application's specific demands. In high-stakes environments like autonomous driving and medical diagnosis, robustness-focused techniques such as VRCP and adversarial training are preferred despite higher computational costs. Resource-constrained situations, like on mobile devices or in real-time systems, benefit from methods like CDF-aware smoothing, which balance efficiency with robustness. For general consumer applications such as recommendation systems and image recognition, Adversarially Robust Conformal Prediction provides moderate robustness and efficient performance. Each approach caters to distinct needs, ensuring optimal performance across varied domains and operational constraints.

VII. CONCLUSION

The evaluation of conformal prediction methods under adversarial conditions highlights various strategies and their respective advantages and challenges. [3] combines randomized smoothing with conformal prediction, striking a balance between robustness and computational demands, thereby enhancing its applicability across a wide range of practical scenarios. Following this, [4] emphasizes the importance of adversarial training for boosting robustness, despite its significant computational costs. [5], on the other hand, relies on neural network verification to provide solid theoretical guarantees, yet faces issues related to complexity and scalability. Lastly, [6] introduces CDF-conscious smoothing, effectively balancing efficiency with robustness, particularly suitable for real-time applications. Each method offers distinct contributions, catering to different aspects of robustness, efficiency, and scalability in adversarial settings.

The current state of conformal prediction in adversarial settings exhibits encouraging progress, despite the existence of formidable challenges. Methods that prioritize robustness often come with high computational expenses, which in turn restrict their scalability and practical applicability. On the other hand, techniques that concentrate on enhancing efficiency might compromise on robustness. Future research should concentrate on diminishing computational complexity, bolstering robustness against a wider spectrum of attacks, and enhancing generalization across various models and datasets. By focusing on these areas, conformal prediction methods can become more adaptable and dependable, thereby making them suitable for an expanded range of real-world applications.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 06, June - 2024

VIII. FUTURE RESEARCH

A. Improving Efficiency

To decrease computational complexity, it is recommended that researchers concentrate on the development of optimization algorithms, such as stochastic gradient descent, utilization of sparse representations, and the utilization of hardware accelerators, including GPUs and TPUs. Additionally, simplifying conformal prediction algorithms and adopting approximate methods may prove beneficial.

B. Expanding Robustness

Improving robustness entails expanding defenses to protect against a broader range of adversarial attacks, such as data poisoning and backdoor attacks. Developing adaptive defense mechanisms and hybrid approaches that integrate multiple defensive strategies can offer comprehensive protection. Furthermore, enhancing robustness verification tools and adversarial example generation methods are essential.

C. Generalization

To enhance generalization across multiple models and datasets, the development of model-agnostic techniques is essential. Strategies such as cross-domain learning, automated hyperparameter tuning, and data augmentation can increase adaptability. Additionally, establishing standardized evaluation
benchmarks and investigating meta-learning for rapid adaptation to new tasks can contribute to improved generalization.
These directions aim to advance the field of conformal prediction by making methods more efficient, robust, and widely applicable across different adversarial scenarios and machine learning models

ACKNOWLEDGMENT

I would like to thank Aakash for their valuable assistance in editing this paper. His attention to detail has helped improve the manuscript.

REFERENCES

[1] V. Vovk, A. Gammerman, and G. Shafer, Algorithmic learning in a random world. 2005. doi: 10.1007/b106715.

[2] C. Szegedy et al., “Intriguing properties of neural networks,” in 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings, 2014.

[3] A. Gendler, T. W. Weng, L. Daniel, and Y. Romano, “Adversarially Robust Conformal Prediction,” in ICLR 2022 - 10th International Conference on Learning Representations, 2022.

[4] Z. Liu et al., “The Pitfalls and Promise of Conformal Inference Under Adversarial Attacks,” arXiv preprint arXiv:2405.08886, 2024.

[5] L. Jeary, T. Kuipers, M. Hosseini, and N. Paoletti, “Verifiably Robust Conformal Prediction,” arXiv preprint arXiv:2405.18942, 2024.

[6] S. H. Zargarbashi, M. S. Akhondzadeh, and A. Bojchevski, “Robust Yet Efficient Conformal Prediction Sets,” in Forty-first International Conference on Machine Learning,

(This work is licensed under a Creative Commons Attribution 4.0 International License.)