What Makes Models Explainable?

Student: Ipek Iscan (i.iscan@student.tudelft.nl), Responsible Professor: Cynthia Liem, Supervisor: Patrick Altmeyer

1. Background

- **Explainable Artificial Intelligence (XAI):** AI systems that can explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future [1].
- **Black-box models:** machine learning algorithms that do not provide any direct explanation for their predictions [2].
- **Counterfactual explanations** work under the premise of strategically perturbing model inputs to understand model behaviour [3].
- **Predictive uncertainty** is defined as the total uncertainty representing the model's confidence in its predictions [4].
- **Adversarial robustness** measures a network's resilience against adversarial attacks [5].
- **Generative capacity** is the system's ability to autonomously generate new, meaningful data points.

2. Research Question

What techniques have been developed to improve the explainability of black-box models, according to existing literature?

1. **Key Techniques:** What are the model improvement techniques used to advance explainability in black-box models?
2. **Impact on Explanation Quality:** How do these techniques impact the explainability of the models?
3. **Predictive Uncertainty:** How do these techniques affect the predictive uncertainty of the models?
4. **Adversarial Robustness:** How do these techniques affect the adversarial robustness of the models?
5. **Generative Capacity:** How do these techniques affect the generative capacity of the models?

3. Methodology

The report was structured according to Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) [6] guidelines.
The snowballing methodology is used as a search approach for the systematic literature review. It is chosen for its effectiveness in uncovering comprehensive and significant studies by iteratively expanding the set of reviewed papers [7].

- **Initial Set Formation:**
  - The initial set of papers was collected using keywords and synonyms from the research question: "black-box model", "explainable", "improve".
  - Engines Used: IEEExplore, Scopus, ArXiv, ACM Digital Library.
  - A date filter was applied to include papers published from 2013 onwards.
- **Snowballing Procedure:**
  - Backward Snowballing: References of selected papers were reviewed to identify new relevant papers.
  - Forward Snowballing: New papers that cited the selected papers were identified.
  - This process was repeated for four iterations to ensure comprehensive coverage.
- **Inclusion and Exclusion Criteria:**
  - Inclusion criteria ensured the review focused on papers introducing techniques to improve model explainability within the Computer Science or Machine Learning fields.
  - Exclusion criteria filtered out papers not written in English, those that only compared XAI methods or explained specific XAI methods, and those with insufficient or unclear methodology details.

4. Results & Findings

The systematic review started with an initial set of 13 papers. Through four iterations of snowballing, additional relevant papers were identified and included. The first iteration added 21 papers, the second iteration contributed 10 more, the third iteration added 4, and the fourth iteration included the final 2 papers, bringing the total to 50 papers. These papers were systematically reviewed, providing a broad foundation for analyzing techniques.

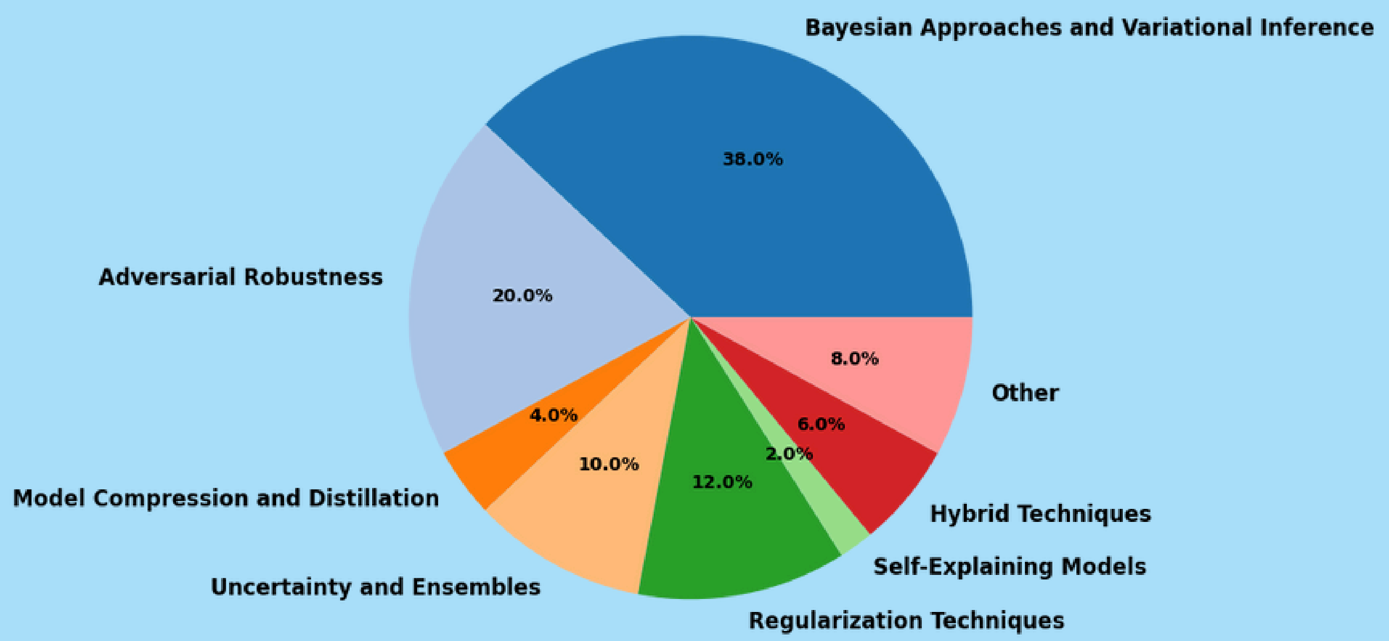

Figure 1: Frequency of Techniques

Figure 1 shows the distribution of different techniques used in the reviewed papers. The motivation behind this graph is to identify which are most commonly employed, providing insights into current trends and potential gaps.

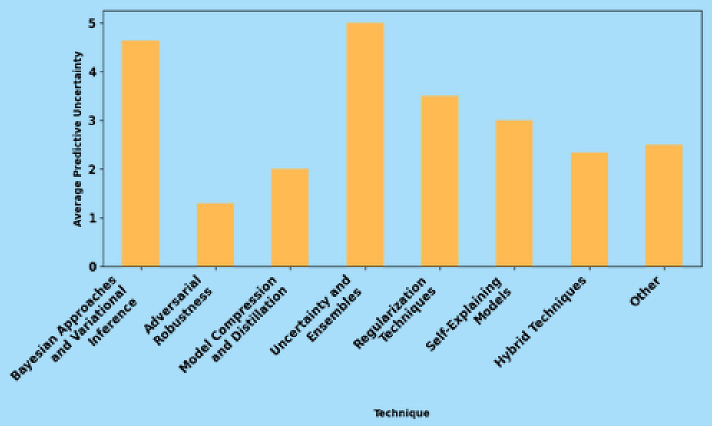

Figure 2: Impact on Predictive Uncertainty

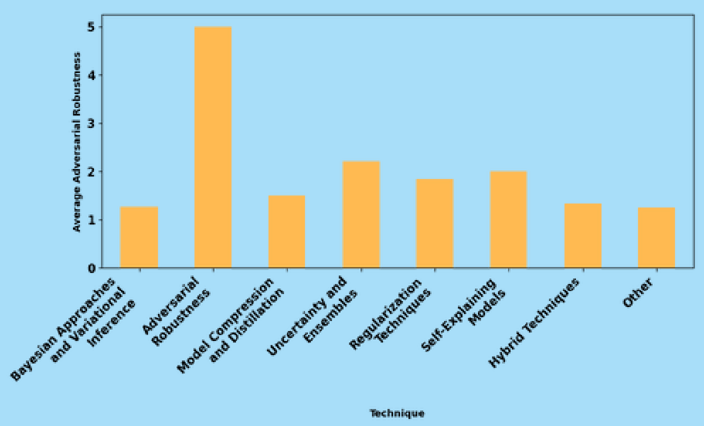

Figure 3: Impact on Adversarial Robustness

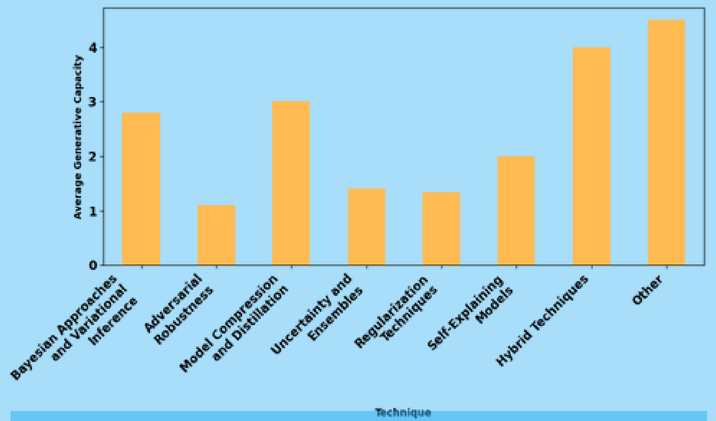

Figure 4: Impact on Generative Capacity

Figures 2, 3, and 4 illustrate the average scores of various techniques on predictive uncertainty, adversarial robustness, and generative capacity, respectively.

5. Conclusion

- **Improvement Techniques:** Various techniques such as Bayesian approaches, adversarial robustness, model compression and distillation, uncertainty and ensembles, regularization, self-explaining models, and hybrid techniques significantly enhance the explainability of black-box models.
- **Impact Analysis:** Uncertainty and ensemble techniques positively impact predictive uncertainty, adversarial training techniques have the greatest impact on adversarial robustness, and hybrid and other innovative techniques enhance generative capacity, contributing to overall explainability.
- **Key Findings:** Bayesian approaches and variational inference were found to be the most common techniques, providing robust probabilistic interpretations that enhance model explainability.
- **Practical Implications:** The practical application of these techniques helps in making AI models more explainable, which can contribute to the quality of counterfactual explanations .

6. Limitations

- **Single Researcher Bias:** The review was conducted by a single researcher, which may introduce bias and errors in paper selection and data extraction.
- **Time and Database Constraints:** The initial set of papers was restricted to those published from 2013 onwards and retrieved from only four databases, potentially missing foundational studies and other relevant sources.
- **Interdisciplinary Knowledge Gaps:** The research required a comprehensive understanding of both machine learning techniques and explainability methods, which might not be fully covered by the primary researcher's expertise.
- **Terminology Assumption:** Using the terms 'interpretability' and 'explainability' interchangeably may obscure important differences in their precise definitions and applications.
- **Scalability and Standardization Issues:** The techniques reviewed may not scale well to larger, more complex models, and there is a lack of standardized metrics to evaluate explainability, making comparisons difficult.

7. Future Work

- **Scalability and Applicability:** Focus on the scalability of these techniques to larger and more complex models, and their applicability across different domains and model architectures.
- **New Techniques Development:** Develop new techniques that advance model explainability.
- **Formally Defined Taxonomy:** Establish a more formally defined taxonomy of techniques. This taxonomy should categorize techniques based on clear and consistent criteria, facilitating better comparison and integration.
- **Feasibility Filter-Free Research:** Conduct future research without feasibility filters to ensure a more comprehensive and unbiased review of the literature.

References

[1] Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. AI Magazine, 40(2), 44-58. https://doi.org/10.1609/aimag.v40i2.2850
[2] Casalicchio, G., Molnar, C., Bischl, B. (2019). Visualizing the Feature Importance for Black Box Models. In: Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N., Ifrim, G. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2018. Lecture Notes in Computer Science(), vol 11051. Springer, Cham. https://doi.org/10.1007/978-3-030-10925-7_40
[3] Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2017. "Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR." Harv. JL & Tech. 31: 841. https://doi.org/10.2139/ssrn.3063289.
[4] Zhang, Z., Wang, S., Hu, S., Liu, F., Hua, X.-S., & Liu, W. (2021). A Simple Baseline for Bayesian Uncertainty in Deep Learning. arXiv. https://arxiv.org/abs/2110.06427
[5] Katz, G., Barrett, C., Dill, D. L., Julian, K., & Kochenderfer, M. J. (2017). Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks. arXiv. https://arxiv.org/abs/1709.02802
[6] Page M J, McKenzie J E, Bossuyt P M, Boutron I, Hoffmann T C, Mulrow C D et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews BMJ 2021; 372 :n71 doi:10.1136/bmj.n71
[7] Wohlin, C. (2014). Guidelines for Snowballing in Systematic Literature Studies and a Replication in Software Engineering. Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE '14). ACM, New York, NY, USA, Article 38, 10 pages. DOI: 10.1145/2601248.2601268