Advancing Explainability in Black-Box Models

A Systematic Literature Review

Ipek Iscan[1]
Supervisors: Cynthia Liem[1]**, Patrick Altmeyer**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Ipek Iscan
Final project course: CSE3000 Research Project
Thesis committee: Cynthia Liem, Patrick Altmeyer, Bernd Dudzik

Abstract

In recent years, the need for explainable artificial intelligence (XAI) has become increasingly important as complex black-box models are used in critical applications. While many methods have been developed to interpret these models, there is also potential in enhancing the models themselves to improve their inherent explainability. This paper investigates various techniques aimed at improving the explainability of black-box models. Through a systematic literature review, these techniques are categorized, and their impact on predictive uncertainty, adversarial robustness, and generative capacity is analyzed to understand how these factors contribute to the overall explainability. The snowballing methodology is used for the systematic literature review, starting with papers retrieved from four databases: IEEExplore, Scopus, ArXiv, and the ACM Digital Library to form the initial set. This process continued with backward and forward snowballing through four iterations, resulting in a total of 50 papers reviewed. Only papers focused on improving model explainability are included in the review. Due to time limitations, additional search constraints are applied for feasibility. The initial set of papers is filtered to those published since 2013. These constraints and their possible impacts are considered when interpreting the results. Findings reveal that techniques such as Bayesian approaches and variational inference, adversarial robustness, model compression and distillation, uncertainty and ensembles, regularization, self-explaining models, and hybrid techniques are used for advancing model explainability. The paper concludes with a discussion on the implications of these techniques for future research.

1 Introduction

The rapid advancement of artificial intelligence (AI) has led to the widespread deployment of complex black-box models in various critical domains such as banking, e-commerce, healthcare, and public services and safety [29]. While these models exhibit high accuracy, they often lack transparency. This opacity poses significant challenges, including difficulties in interpreting model behaviors [42], a lack of trust from end-users [59], and issues in meeting regulatory requirements [27].

Related Work

In response to the challenges posed by the opacity of black-box models, the field of explainable AI (XAI) has emerged, focusing on developing methods to make AI models more interpretable. Traditional XAI approaches, such as SHAP [44] and LIME [59], provide post-hoc explanations for model predictions. These methods are popular due to their ability to attribute importance to individual features [51] and generate understandable explanations for model outputs [5]. However, they do not address the fundamental opacity of the models themselves, as they offer explanations that are separate from the model's inherent structure and processes [60].

Counterfactual explanations (CE) also provide post-hoc insights but do so by identifying changes in input data that would lead to different outputs, thereby offering a more intuitive way to understand model decisions. CE ensures full fidelity to the model by construction, meaning that the generated counterfactuals faithfully represent the model's decision-making process [2]. This characteristic makes CE particularly useful for generating explanations that are accurate and reflective of the model's internal logic.

Recent advancements in CE, such as the ECCCo framework (Explainable Counterfactual Explanations through Constrained Optimization), further enhance this approach. The ECCCo framework generates counterfactual explanations that are as plausible as the model

allows, providing a more faithful representation of the model's decision boundaries [1]. It helps in evaluating the inherent explainability of the models.

Despite the usefulness of XAI methods, there is an interest in enhancing the inherent explainability of black-box. This research aims to investigate what work has been done to improve black-box models to deliver better explanations. Unlike other research that generally provides theoretical insights into the challenges and future directions of explainable AI [61], this study focuses less on comparing XAI methods and focuses more on "Provided we have tools to faithfully explain models, how do some models deliver better explanations?". It offers insights into what makes models explainable, aiming to advance the understanding of the factors contributing to the explainability of black-box models. Focusing on these improvements helps to identify which models produce more plausible and faithful counterfactual explanations.

Research Questions

Understanding and improving the inherent explainability of black-box models involves examining various aspects that contribute to model transparency and trustworthiness [62]. Analyzing predictive uncertainty is crucial because it provides insights into how confident a model is in its predictions. Also, it has been shown that uncertainty quantification methods helps reduce the complexity of explanations [20]. When models can quantify their uncertainty, users can better trust and interpret the decisions made by the model, as they have a clearer understanding of the model's confidence in different scenarios [49].

Adversarial robustness is another critical aspect because models that are resistant to adversarial attacks are less likely to be misled by small, malicious perturbations. This stability ensures that the explanations provided by the model remain valid and reliable even under adversarial conditions [47]. It has been shown that adversarial robustness improves explainability [6].

Lastly, generative capacity is essential for producing realistic counterfactual explanations and for making the latent representations of data more interpretable [67]. Generative models that accurately capture the data distribution help generate more plausible and faithful counterfactual explanations, which in turn enhance the model's explainability by providing a clearer understanding of the data generation process and hypothetical scenarios [39]. This report analyzes how these aspects in various techniques improve the explainability of black-box models.

This approach aligns with the overarching goal of advancing XAI by not just explaining the models by various methods, but by enhancing the models to inherently support better explanations. This research aims to systematically review and categorize the various techniques developed to enhance the explainability of black-box models. The aim is to understand how these techniques impact the model's ability to provide advanced explanations. Sometimes, both interpretability and explainability terms can be used interchangeably in the broad general sense of understandability in human terms [14], and this is also the case in this paper. The possible consequences of this assumption are considered in Chapter 4.

The main contributions of this research are as follows:

- **C1:** Categorization of techniques aimed at improving the inherent explainability of black-box models.
- **C2:** An analysis of the impact of these techniques on predictive uncertainty, adversarial robustness, and generative capacity.
- **C3:** Insights into the practical implications of these techniques for enhancing model explainability.

While trying to answer the main question, the sub-questions in Table 1 are evaluated. These sub-questions give insight on what the overall research goals seek to achieve. Detailed reasoning for why these questions were addressed can be found in the 'Motivation' column. C1, C2 and C3 refer to contributions.

Table 1: Sub-questions to help answer the main research question with their motivations

	Sub-question	**Motivation**
1	What are the key model improvement techniques used to enhance explainability in black-box models?	Related to C1. Identifying the various techniques allows for a clear categorization of the different approaches researchers have taken. This helps in understanding the diversity of methods and their respective contributions.
2	How do these model improvement techniques impact the quality of explanations provided by the models?	Related to C1 and C3. Assessing the impact on explanation quality ensures that the focus remains on improving interpretability rather than just performance metrics. This sub-question evaluates how effective these techniques are in making the models more explainable.
3	How do these techniques affect the predictive uncertainty?	Related to C2. Understanding this impact helps to determine if improved explainability also leads to better handling of uncertainty in predictions, which is crucial for trust and reliability in real-world applications.
4	How do these techniques affect the adversarial robustness of the models?	Related to C2. Evaluating the impact on adversarial robustness ensures that the models not only provide better explanations but are also resilient to adversarial attacks. This is important for the practical deployment of these models in security-sensitive applications.
5	How do these techniques affect the generative capacity of the models?	Related to C2. Investigating this relationship helps to understand if improved explainability compromises the model's ability to generate new data or representations, which is essential for tasks like data augmentation and anomaly detection.
6	What are the key findings and insights from the research on improving black-box models for better explanations?	Related to C3. Summarizing the key findings provides a concise overview of the major contributions and insights gained from the research, highlighting the most effective and innovative approaches.

The rest of the paper is organized as follows: Section 2 describes the methodology used, including the initial sourcing from four databases and the subsequent snowballing methodology. Section 3 presents the results of the research. Section 4 discusses the reproducibility and ethical aspects of the review. Section 5 discusses the results, and Section 6 concludes with recommendations for future work.

2 Methodology

This paper is structured according to Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines [56]. This section outlines the methodology employed

to conduct a systematic literature review using the snowballing approach. The aim is to provide a clear and detailed explanation of the steps taken to answer the research questions and to justify the chosen methods.

2.1 Snowballing Procedure

In this paper, the snowballing methodology is used as a search approach for the systematic literature review. Snowballing is chosen for its effectiveness in uncovering comprehensive and significant studies by iteratively expanding the set of reviewed papers. Guidelines from Wohlin [71] are followed for a systematic snowballing procedure. The procedure consists of two main phases: the initial sourcing of papers to form a start set and the subsequent snowballing iterations.

2.1.1 Start Set

The first step in the snowballing process was to establish a start set of papers. The criteria for selecting relevant papers for the start set included:

- Identifying records from 4 different databases to avoid missing relevant papers from independent clusters.
- Including a sufficient number of papers to ensure a diversity of techniques, preventing all papers from focusing on the same technique.
- Diversity in terms of publishers, years, and authors to capture a wide range of perspectives.

To form a start set, relevant keywords and their synonyms from the research question listed in Table 2 were used:

Table 2: Relevant keywords and their synonyms from the research question

Keywords	**Synonyms**
black-box model	"black-box model*", "opaque model*", "complex model*", "neural network*"
explainable	explainab*, interpret*, "model explanation", clear*
improve	improv*, enhanc*, advanc*, refin*, optimiz*

The papers in the initial set are collected from the following databases with a date filtering to include papers from 2013 onwards as follows: IEEE Xplore[1] ([12], [74]); ACM Digital Library[2] ([3], [8], [12], [16], [24], [58], [65]); ArXiv[3] ([3], [8], [16], [24], [26], [30], [32], [39], [47], [58], [74]); Scopus[4] ([3], [24], [26], [30], [39], [47], [58], [74]). After removing duplicated papers, 13 papers were selected for the initial set: [3], [8], [12], [16], [24], [26], [30], [32], [39], [47], [58], [65], [74].

2.1.2 Iterations

After establishing the start set, the snowballing process proceeded through four iterations, each consisting of backward and forward snowballing.

[1] `https://ieeexplore.ieee.org`
[2] `https://dl.acm.org`
[3] `https://arxiv.org/`
[4] `https://www.scopus.com`

Backward Snowballing: The references of the selected papers were reviewed to identify new relevant papers.

Forward Snowballing: Involved identifying new papers that cited the selected papers. For both backward and forward snowballing, the following steps were taken:

1. Initial Screening: The titles were examined to determine their relevance.
2. Context Evaluation: The context in which the paper was cited was considered to evaluate its relevance.
3. Abstract Review: The abstracts of the potential papers were read, followed by selective reading of other parts if necessary.
4. Inclusion Decision: Only papers that met the inclusion criteria from Section 2.1.3 were included for further snowballing.

The snowballing process is illustrated in Figure 1. It includes the results of the four iterations (no more papers were found after iteration 4).

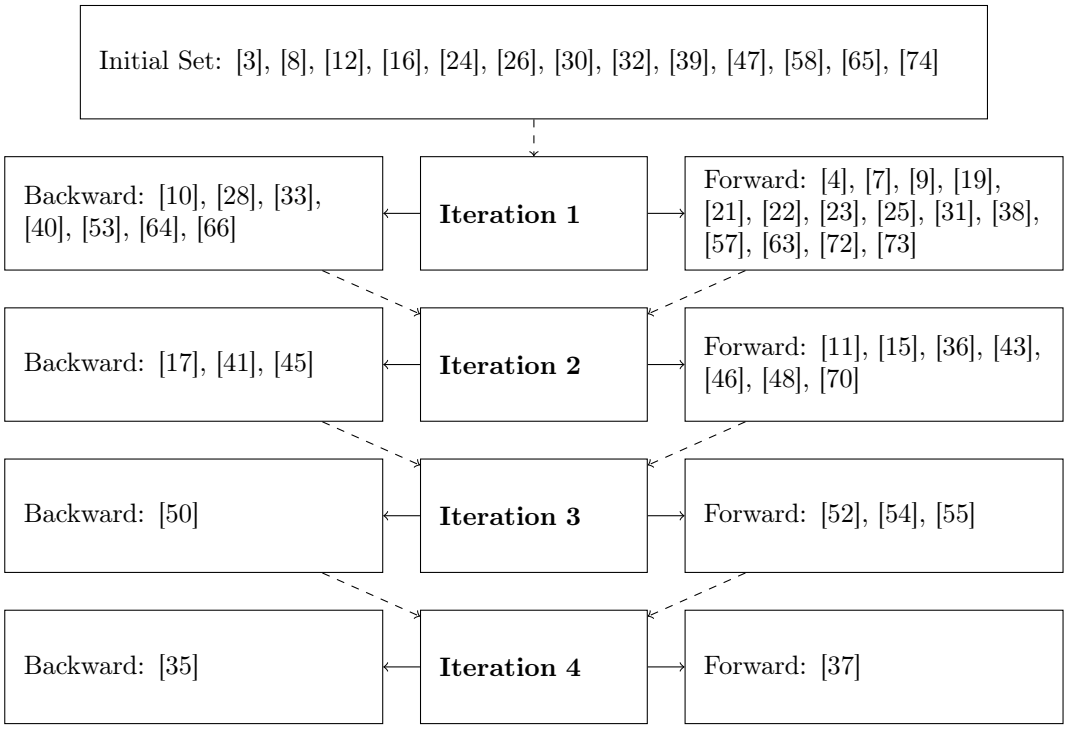

Figure 1: The Snowballing Process

2.1.3 Inclusion and Exclusion Criteria

To ensure the quality and relevance of the papers included in the literature review, the following inclusion and exclusion criteria were applied:

Inclusion Criteria:

- Paper introduces a technique to improve model explainability. (*Ensuring relevance to the core objective of advancing explainability*)

- Paper introduces a framework to improve the model which can directly/indirectly advance explainability. (*Including frameworks that contribute to the overall goal of better model explainability*)
- Paper is from the Computer Science or Machine Learning field. (*Maintaining focus on relevant academic disciplines*)

Exclusion Criteria:

- Paper not written in English. (*Ensuring accessibility and ease of review*)
- Paper only comparing XAI methods. (*Focusing more on model improvements rather than comparisons*)
- Paper only using existing XAI methods on various models. (*Focusing more on model improvements rather than explanatory descriptions*)
- Paper only explaining how a specific XAI method works. (*Focusing more on model improvements rather than XAI methods*)
- Paper with insufficient or unclear methodology details. (*Ensuring rigor findings*)

2.1.4 Data Extraction

This paper aims to answer "What makes models more explainable?". Many papers exist to compare different explainability methods [34], but there is currently a gap in the literature when it comes to reviewing the improvements done to black-box models to advance explainability. To answer our questions, papers are reviewed and data is extracted using specific criteria. The extracted information includes the title, author(s), and year of publication to prevent duplication, considering that different databases might include the same papers and many papers are frequently referenced or cited by others. The data extraction focused on answering specific sub-questions: the model improvement technique (Sub-question 1), the impact on explanation quality (Sub-question 2), the relationship with predictive uncertainty (Sub-question 3), the relationship with adversarial robustness (Sub-question 4), the relationship with generative capacity (Sub-question 5), and key findings (Sub-question 6). The results of this data extraction are discussed in Section 3.

2.1.5 Reliability and Validity

The reliability and validity of the snowballing procedure were ensured by adhering to systematic guidelines. Specifically, the guidelines from Wohlin [71] were followed to maintain consistency throughout the iterations. The guidelines followed included: *Systematic Search Process* to ensure a structured approach for both backward and forward snowballing, *Consistent Application of Inclusion and Exclusion Criteria* throughout the iterations, *Documentation of Decisions* by carefully documenting included papers and *Contextual Evaluation* where the context is evaluated in which references and citations occurred to ensure relevance. Each step of the snowballing process was carefully documented, and all decisions regarding the inclusion or exclusion of papers were based on predefined criteria from Section 2.1.3.

By iteratively expanding the set of reviewed papers through backward and forward snowballing, the study achieved a thorough understanding of the research area, uncovering various studies that might have been missed using traditional search methods alone.

3 Results

This section presents the analysis of the literature reviewed on improving black-box models to advance explainability. The data collected has been discussed in several subsections, including model improvement techniques, their impact on explanation quality, and their influences on predictive uncertainty, adversarial robustness, and generative capacity. Each subsection delves into these aspects in detail, providing an overview of the state of research in this domain.

3.1 Model Improvement Techniques

This section categorizes and summarizes the techniques used in the selected papers to improve the explainability of black-box models. Further details about the categorization can be found in Appendix A.

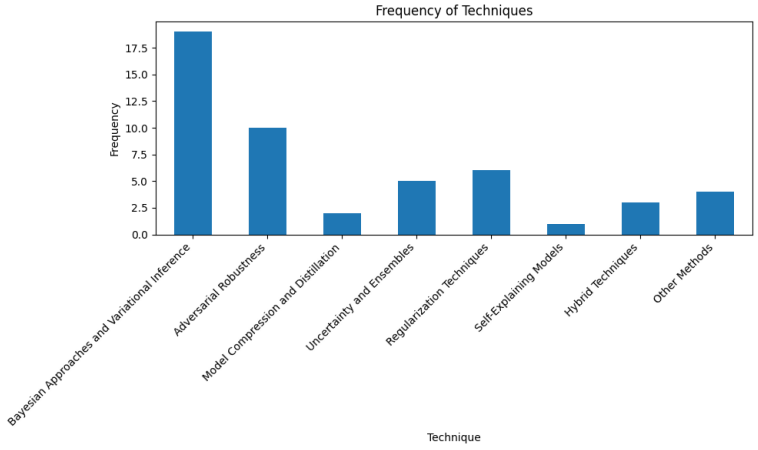

Figure 2: Frequency of Techniques

The bar chart in Figure 2 shows the distribution of different techniques. This visualization helps in understanding the frequency of use of different methods. The most common technique in this review was Bayesian approaches and variational inference with 19 relevant papers and the least common was self-explaining methods with 1 relevant paper.

Bayesian approaches and variational inference techniques integrate probabilistic models into neural networks, providing a framework for handling uncertainty in model predictions. Models like Bayesian Neural Networks (BNNs) and Variational Autoencoders (VAEs) use Bayesian inference to estimate the distribution over model parameters or latent variables. This allows the model to express uncertainty about its predictions and to learn representations that align closely with the data distribution. Methods such as dropout as a Bayesian approximation, which treats dropout as a variational Bayesian method to estimate uncertainty in model predictions [24], Bayes-by-Backprop, which performs variational inference by learning distributions over network weights [8], and Auto-Encoding Variational Bayes, which learns latent variable models for representing complex data distributions [39] are further refinements. The relevant 19 papers for this category are: [7, 8, 9, 22, 23, 24, 25, 28, 31, 35, 36, 39, 45, 46, 52, 53, 54, 58, 74].

Adversarial robustness techniques are designed to make models resistant to adversarial attacks, which are small perturbations in input data that can cause significant changes in model output. Methods such as adversarial training, which incorporates adversarial examples into the training process [47], and provable defenses that use mathematical techniques to guarantee robustness against certain types of attacks [72], are commonly employed. These techniques enhance the stability of the model by ensuring that its predictions remain consistent even when the inputs are slightly altered in a way designed to deceive the model. The relevant 10 papers for this category are: [4, 12, 26, 40, 47, 57, 63, 66, 72, 73].

Model compression and distillation techniques aim to reduce the complexity of neural networks while maintaining their performance. Model distillation [32] involves training a smaller "student" model to mimic the behavior of a larger "teacher" model [69]. Compression techniques, such as pruning, quantization, and low-rank factorization, reduce the number of parameters and operations in the model [18]. These methods streamline the model by eliminating redundancies and focusing on the most critical parts of the network, thereby simplifying its structure without significant loss in accuracy. The relevant 2 papers for this category are: [10, 32].

Uncertainty and Ensembles represent another important technique used to enhance model explainability. Ensemble methods combine multiple models to improve prediction accuracy and estimate uncertainty. Techniques like bagging, boosting, and stacking involve training several models and aggregating their predictions [41]. Bayesian ensembling combines the strengths of Bayesian inference with ensemble methods to provide robust uncertainty estimates [55]. By leveraging the diversity among the ensemble members, these methods capture a broader range of model behaviors and uncertainties, resulting in more reliable predictions. The relevant 5 papers for this category are: [21, 41, 48, 55, 70].

Regularization techniques aim to prevent overfitting and improve the generalization of neural networks. Techniques such as L0/L1/L2 regularization, dropout, and batch normalization add constraints or modifications to the training process to ensure that the model does not become overly complex. These techniques penalize large weights, enforce sparsity, or introduce noise during training, encouraging the model to learn more robust and generalized patterns rather than memorizing the training data [68]. The relevant 6 papers for this category are: [33, 37, 43, 50, 64, 65].

Self-explaining models are designed with inherent mechanisms that provide explanations as part of their architecture. These models integrate explanation-generation processes directly into their structure, ensuring that the outputs are accompanied by understandable reasons. Examples include attention mechanisms in neural networks [3], which highlight the parts of the input that are most relevant to the prediction, and models that output human-readable rules or decision trees. The relevant paper for this category is: [3].

Hybrid techniques combine multiple methods to enhance both model performance and explainability. These approaches integrate various improvement techniques, such as combining Bayesian inference with adversarial training or using ensemble methods alongside regularization. The goal is to leverage the strengths of different techniques to create models that are both robust and interpretable [38]. The relevant 3 papers for this category are: [17, 30, 38].

Other techniques include various innovative techniques that contribute to model explainability but do not fit into the previously mentioned categories. Techniques like neural ordinary differential equations (ODEs) provide a mathematical framework for modeling continuous-time dynamics, enhancing explainability by allowing users to track how inputs transform continuously through the model [16]. Augmented neural ODEs expand this con-

cept by adding latent dimensions to better capture complex data distributions, resulting in more flexible modeling and clearer explanations of the model's behavior [19]. Additionally, the exploration of disentangled representations focuses on learning representations that capture distinct, interpretable features of the data [11]. By identifying and isolating factors contributing to disentanglement in variational autoencoders, these techniques help in understanding the underlying data generation processes, thus improving model interpretability [15]. These methods enhance the transparency and interpretability of complex models by leveraging advanced mathematical frameworks and representation learning techniques. The relevant 4 papers for this category are: [11, 15, 16, 19].

3.2 Impact on Explanation Quality

The improvement techniques directly impact the quality of explanations provided by the models by making their decision-making processes more transparent and understandable.

Bayesian Approaches and Variational Inference
The integration of Bayesian methods significantly enhances the explainability of black-box models by providing probabilistic interpretations of predictions. This means that instead of giving a single deterministic output, the model provides a distribution, which reflects its confidence in the predictions. This helps users understand not just what the model predicts, but how certain it is about those predictions. This approach is demonstrated in techniques such as Bayesian Convolutional Neural Networks [23], where uncertainty estimation helps in understanding the model's confidence. In safety-critical applications, knowing the uncertainty of a prediction can be crucial. By capturing and communicating this uncertainty, Bayesian models make their decision-making process more transparent, thereby increasing trust and enabling better-informed decisions based on model outputs [54].

Adversarial Robustness
Improving adversarial robustness contributes to the reliability and trustworthiness of model explanations. When a model is robust to adversarial attacks, its predictions are less likely to be affected by small changes in input data. This stability ensures that the explanations generated by the model are consistent and reliable. Robust models provide explanations that are more aligned with the actual decision boundaries learned from the training data, making it easier to understand and trust the modelâs behavior [13]. This increased reliability is crucial for applications where the consequences of incorrect predictions are severe.

Model Compression and Distillation
These techniques enhance explainability by reducing model complexity while retaining predictive accuracy. Training a single model to mimic the behavior of an ensemble captures the ensemble's decision boundaries in a more interpretable form [10]. Similarly, using soft targets provided by a larger model during knowledge distillation allows the distilled model to retain essential predictive features with fewer parameters [32]. This process improves interpretability by simplifying the model structure and making it easier to trace and understand the decision-making process, while also being efficient for deployment in resource-constrained environments.

Uncertainty and Ensembles
Ensembles enhance the explainability of models by providing multiple perspectives on the

data and offering insights into the variability of predictions. The aggregated output from an ensemble reflects a consensus view, which tends to be more robust and less sensitive to individual model biases [21]. Additionally, ensembles can provide confidence intervals around predictions, helping users understand the range of possible outcomes and the confidence of the model in its predictions. This added layer of transparency is crucial for building trust in the model, as it communicates the reliability and stability of the predictions.

Regularization Techniques
Regularization enhances the explainability of models by simplifying their structure and making their decision boundaries smoother and more generalizable. A regularized model is less likely to overfit to noise in the training data [43], leading to more consistent and reliable predictions. This consistency translates to clearer and more straightforward explanations, as the model's behavior is governed by broader, more interpretable patterns. Regularized models are easier to analyze and understand, as they tend to avoid the complexities and intricacies of overfitted models [50].

Self-Explaining Models
Self-explaining models enhance explainability by integrating interpretability directly into their architecture. Self-Explaining Neural Networks (SENNs) decompose predictions into concept vectors and corresponding importance scores, making it clear which features contribute to decisions [3]. This method ensures that users can understand the rationale behind each decision, which is particularly useful in critical applications.

Hybrid Techniques
Hybrid techniques enhance explainability by integrating multiple methods. InfoGAN combines generative adversarial networks with information maximization to learn interpretable representations by maximizing mutual information between latent variables and generated data [17]. Generating visual explanations using natural language justifications helps users understand predictions by combining classification and sentence generation, improving explanation quality with discriminative class label loss during training [30]. Additionally, semi-supervised learning with deep generative models captures data distributions and explains underlying data structures by combining generative modeling with semi-supervised learning techniques [38]. This approach provides robust and comprehensive explanations, making it easier for users to trust and comprehend the model's behavior in complex scenarios.

Other Techniques
Other techniques enhance explainability by introducing novel ways of understanding and interpreting the modelâs internal processes. Techniques like disentangled representations allow for a clearer interpretation of the latent space, making it easier to understand how different factors influence the modelâs outputs [11]. Neural ODEs provide a mathematically rigorous framework that enhances transparency and interpretability by modeling the continuous evolution of data [19]. These innovative approaches contribute to a deeper and more nuanced understanding of model behavior, expanding the toolkit available for making black-box models more transparent and explainable.

3.3 Impact on Predictive Uncertainty

Predictive uncertainty plays a role in model explainability. Techniques that quantify uncertainty help in understanding and trusting model predictions. Appendix B represents the scores for each paper from 1 to 5 to represent their impact on predictive uncertainty. Based on their model improvement technique, these papers are clustered and Figure 3 shows the average values for each technique.

The highest correlation with predictive uncertainty is shown by Uncertainty and Ensembles, with the highest average score. For example, in [41], it is demonstrated how combining multiple models in an ensemble can provide robust uncertainty estimates. This is achieved by aggregating the predictions from different models, which helps in capturing the variability and confidence in the predictions.

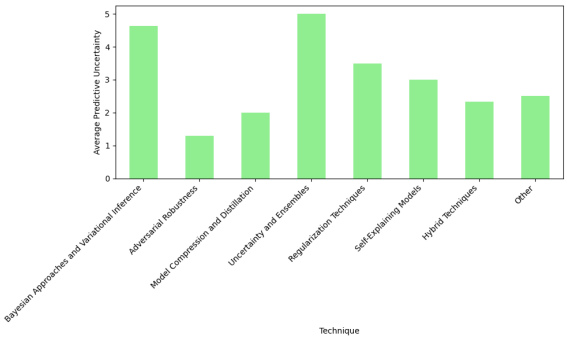

Figure 3: *Impact on Predictive Uncertainty*

Bayesian approaches and variational inference follow closely. Techniques like [24] and [23] provide a probabilistic framework for handling uncertainty in model predictions. By treating model parameters as random variables and estimating their distributions, these methods offer insights into the confidence of the modelâs predictions.

Regularization techniques, such as those discussed in [64], show moderate correlations with predictive uncertainty. These techniques add noise during training to prevent overfitting, indirectly helping in understanding the confidence of the modelâs predictions.

Self-explaining models and hybrid techniques exhibit moderate correlations as well. Self-explaining models like those described in [3] provide inherent explanations for their predictions, enhancing the understanding of model confidence. Hybrid techniques, combining Bayesian methods with regularization, further enhance the reliability of predictions.

Model compression and distillation, along with other techniques, show lower correlations. Techniques like [32] focus more on reducing model complexity while maintaining performance, with less emphasis on directly addressing predictive uncertainty.

3.4 Impact on Adversarial Robustness

Adversarial robustness is important for ensuring that model explanations are reliable and not easily manipulated. Appendix B includes the scores for each paper from 1 to 5 to represent their impact on adversarial robustness. Based on their model improvement technique, these papers are clustered and Figure 4 shows the average values for each technique.

Techniques such as those discussed in [47] use adversarial training, incorporating adversarial examples into the training process to improve robustness. Uncertainty and ensembles techniques also demonstrate a high correlation with adversarial robustness. These methods, such as those in [41] leverage the diversity among multiple models to improve robustness against adversarial attacks by providing more stable and reliable predictions.

Self-explaining models exhibit a moderate correlation with adversarial robustness. Self-explaining models, like those in [3] provide explanations that are inherently robust to ad-

versarial perturbations by integrating interpretability directly into their architecture.

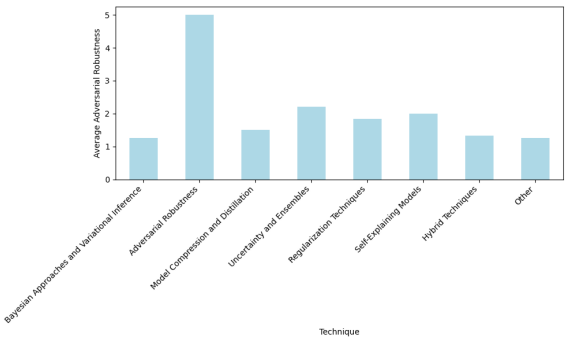

Figure 4: *Impact on Adversarial Robustness*

Regularization techniques show a significant correlation with adversarial robustness. Methods like those discussed in [33] help make models robust to overfitting and adversarial attacks by enforcing constraints on the model parameters.

Model compression and distillation techniques also show a notable correlation with adversarial robustness. Techniques such as those described in [32] use model distillation to improve robustness by training a smaller, simpler model to mimic the behavior of a more complex one, stabilizing the model's responses to adversarial inputs.

Hybrid techniques demonstrate a lower impact on adversarial robustness. These methods, such as those in [4] combine adversarial training with other techniques to enhance robustness, highlighting the importance of a multifaceted approach.

Bayesian approaches and variational inference and other methods show lower correlations with adversarial robustness. These techniques, while excelling in areas like uncertainty estimation and generative capacity, are not primarily designed for enhancing robustness against adversarial attacks.

3.5 Impact on Generative Capacity

Generative models can provide insights into the data generation process, contributing to model explainability. Appendix B represents the scores for each paper from 1 to 5 to represent their impact on generative capacity. Based on their model improvement technique, these papers are clustered and Figure 5 shows the average values for each technique.

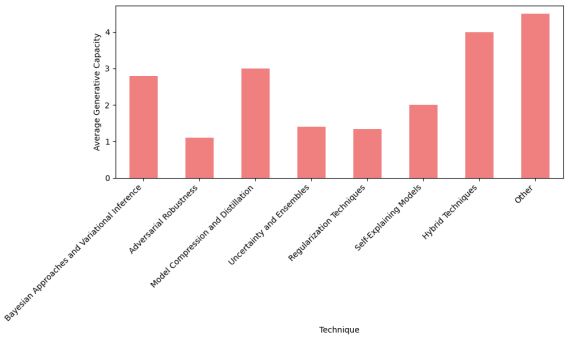

Figure 5: *Impact on Generative Capacity*

"Other" techniques and "Hybrid Techniques" show the highest correlation with generative capacity, reflecting their focus on advanced generative processes. For instance, techniques in [11] aim to disentangle the latent space, enhancing the interpretability and generative capabilities of models.

Model compression and distillation shows a relatively high correlation with generative capacity. Techniques such as [32] help streamline the model, making it easier to generate new data representations while retaining the original model's knowledge. Bayesian Approaches and Variational Inference also exhibit a high correlation with generative capacity. Techniques like [39] capture the data distribution and generate realistic samples, crucial

for meaningful counterfactual explanations. Similarly, [58] demonstrates how integrating Bayesian methods can improve generative capacity by modeling the underlying data distribution.

Uncertainty and ensembles, regularization techniques, and self-explaining models show moderate correlations with generative capacity. Methods like [38] leverage generative models to improve interpretability by capturing and explaining the underlying data distribution.

Adversarial robustness shows the least correlation with generative capacity, indicating that its primary focus is on enhancing model stability and robustness rather than on data generation. Techniques in this category prioritize preventing models from being misled by adversarial inputs, which does not directly contribute to their generative capabilities.

4 Responsible Research

This section mentions the risk of bias in this research, reproducibility of results and further ethical considerations for a responsible research.

Risk of Bias
This systematic literature review (SLR) was conducted by a single researcher, which introduces potential risks of bias and errors. The involvement of a single researcher may lead to inadvertent mistakes during the paper selection and data extraction stages. To mitigate these risks, the methodology was carefully standardized, and clear criteria were established for selecting relevant studies and extracting data. However, it is acknowledged that complete elimination of bias is not achievable.

An additional consideration is the interdisciplinary nature of this research, which requires a thorough understanding of both machine learning techniques and methods for improving model explainability. While the primary researcher has a background in Computer Science, efforts were made to bridge any knowledge gaps by consulting relevant literature and seeking guidance from experts in the field. This approach ensured that the interpretation of results and conclusions drawn were based on a comprehensive understanding of the subject matter.

In this literature review 'interpretability' and 'explainability' terms are be used interchangeably. This decision is based on the understanding that both terms, while sometimes nuanced differently in specific research contexts, fundamentally aim to address the same core issue: making AI models and their decisions more understandable to humans. However, it is worth acknowledging that this interchangeability carries a risk. Specifically, conflating the terms may obscure important differences in their precise definitions and applications in some cases.

Reproducibility of Results
To ensure the reproducibility of this review, the methodology was described in detail. The search strategy, including databases and keywords used, was thoroughly documented. The inclusion and exclusion criteria for selecting relevant papers were clearly defined. All relevant stages of the review process, including paper selection and data extraction, were documented. By following the described methodology, future researchers can verify the results and potentially uncover additional insights. All papers included in the review were cited and listed in the bibliography, providing a comprehensive overview of the literature analyzed.

Ethical Considerations
In conducting this research, ethical considerations were taken into account to ensure the in-

tegrity and credibility of the work. All sources were properly cited to give credit to original authors, and efforts were made to avoid plagiarism. The analysis was conducted objectively, and any potential conflicts of interest were disclosed.

By adhering to these principles, this research strives to maintain high standards of ethical conduct and scientific rigor, ensuring that the findings are reliable and contribute meaningfully to the field of explainable AI.

5 Discussion

This section summarizes the findings of this systematic literature review. Before discussing these findings, possible impact of the start set and feasibility constraints are mentioned.

Possible Impact of Start Set and Feasibility Constraints on Results
Identifying a start set of papers is a challenge for the snowballing procedure [71]. Due to time constraints, the initial set of papers was restricted to those published from 2013 onwards, considering the inclusion of recent advancements but potentially excluding foundational studies and introducing a bias towards current trends. The review initially focused on papers retrieved from four databases: IEEExplore, Scopus, ArXiv, and the ACM Digital Library. This approach helps manage the volume of publications. However, in subsequent iterations, papers from different databases (e.g. Semantic Scholar [5] [28]) and earlier years (e.g. 2006 [10]), were considered to broaden the scope and diversity of the review. The quality of the initial papers is critical, as influential papers are likely to lead to other related studies. While the constraints applied to the initial set were necessary for feasibility, they may affect the comprehensiveness and diversity of the review. These impacts were considered to ensure a thorough analysis of techniques aimed at improving the explainability of black-box models.

Results discussion
The analysis of the literature reveals several key findings regarding techniques to enhance the explainability of black-box models. Bayesian approaches and variational inference, identified in 19 papers, improve predictive uncertainty by providing probabilistic interpretations that help users understand the confidence of model predictions [24]. Adversarial robustness techniques, found in 10 papers, enhance the stability of model explanations by making predictions more reliable under adversarial conditions, ensuring consistent explanations even with perturbed inputs [47].

Model compression and distillation techniques, identified in 2 papers, simplify models by reducing complexity and focusing on critical parts of the network, thereby making the models more interpretable [32]. Uncertainty and ensemble methods, discussed in 5 papers, enhance explainability by providing robust uncertainty estimates through the combination of multiple models, capturing variability and confidence in predictions [41]. Regularization techniques, present in 6 papers, prevent overfitting and improve generalization by adding noise during training or enforcing constraints on parameters, leading to more consistent predictions and clearer decision-making processes [64].

Self-explaining models, although represented by only one paper, enhance transparency by integrating explanation mechanisms directly into the model architecture, providing built-in explanations for predictions [3]. Hybrid techniques, discussed in 3 papers, combine multiple methods to leverage their strengths, resulting in models that are both robust and

[5] `https://www.semanticscholar.org/`

interpretable, improving interpretability by capturing and explaining underlying data distributions [38].

Neural ordinary differential equations (ODEs) provide a framework for modeling continuous-time dynamics, enhancing interpretability by allowing users to track how inputs transform continuously through the model [16]. Disentangled representations improve model interpretability by learning distinct, interpretable features of the data, facilitating a better understanding of the underlying data generation processes [11].

6 Conclusions and Future Work

This research aimed to answer the question: "What makes models more explainable?". Through the review of 50 related papers, techniques that enhance model explainability are identified and categorized. Findings highlight the importance of Bayesian approaches, adversarial robustness, model compression and distillation, uncertainty and ensembles, regularization techniques, self-explaining models, hybrid techniques, and other innovative methods in making black-box models more explainable.

Improving the inherent explainability of black-box models influence how plausible and faithful counterfactuals are. Counterfactual explanations are post-hoc, but they offer full fidelity by construction [2], ensuring that the generated explanations faithfully represent the model's decision boundaries. Techniques that improve model explainability contribute to the quality of CEs, making them more understandable and reliable.

In conclusion, while there has been significant progress in improving black-box model explainability, ongoing research is crucial for addressing remaining challenges and fully realizing the potential of these techniques in practical applications. Open issues remain, such as the scalability of these techniques to larger and more complex models and the need for standardized metrics to evaluate explainability. Future research should focus on developing new techniques to enhance explainability, and investigating their applicability across different domains and model architectures to provide deeper insights into their generalizability. Combining these methods with counterfactual explanations could create comprehensive frameworks for model explainability. Addressing scalability and applicability across different types of models and datasets is essential to ensure these advancements lead to more trustworthy and explainable AI systems.

References

[1] Patrick Altmeyer, Mojtaba Farmanbar, Arie van Deursen, and Cynthia C. S. Liem. Faithful Model Explanations Through Energy-Constrained Conformal Counterfactuals. 2023.

[2] Patrick Altmeyer, Arie van Deursen, and et al. Explaining Black-Box Models Through Counterfactuals. In *Proceedings of the JuliaCon Conferences*, volume 1, page 130, 2023.

[3] David Alvarez-Melis and Tommi S. Jaakkola. Towards Robust Interpretability with Self-Explaining Neural Networks, 2018.

[4] Maksym Andriushchenko and Nicolas Flammarion. Understanding and improving fast adversarial training. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[5] Alejandro Barredo Arrieta, Natalia Diaz-Rodriguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI, 2019.

[6] Maximilian Augustin, Alexander Meinke, and Matthias Hein. Adversarial Robustness on In- and Out-Distribution Improves Explainability, 2020.

[7] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859â877, April 2017.

[8] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight Uncertainty in Neural Networks, 2015.

[9] Shahin Boluki, Randy Ardywibowo, Siamak Zamani Dadaneh, Mingyuan Zhou, and Xiaoning Qian. Learnable Bernoulli Dropout for Bayesian Deep Learning, 2020.

[10] Cristian Buciluundefined, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 535â541, New York, NY, USA, 2006. Association for Computing Machinery.

[11] Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE, 2018.

[12] Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017.

[13] Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks, 2017.

[14] Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(8), 2019.

[15] Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders, 2019.

[16] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural Ordinary Differential Equations, 2019.

[17] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets, 2016.

[18] Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. A Survey of Model Compression and Acceleration for Deep Neural Networks, 2020.

[19] Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented Neural ODEs, 2019.

[20] Duarte Folgado, Marilia Barandas, Lorenzo Famiglini, Ricardo Santos, Federico Cabitza, and Hugo Gamboa. Explainability meets uncertainty quantification: Insights from feature-based model fusion on multimodal time series. *Information Fusion*, 100:101955, 07 2023.

[21] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep Ensembles: A Loss Landscape Perspective, 2020.

[22] Yarin Gal. Uncertainty in Deep Learning. 2016.

[23] Yarin Gal and Zoubin Ghahramani. Bayesian Convolutional Neural Networks with Bernoulli Approximate Variational Inference, 2016.

[24] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning, 2016.

[25] Yarin Gal, Jiri Hron, and Alex Kendall. Concrete Dropout, 2017.

[26] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples, 2015.

[27] Bryce Goodman and Seth Flaxman. European Union Regulations on Algorithmic Decision Making and a âRight to Explanationâ. *AI Magazine*, 38(3):50â57, September 2017.

[28] Alex Graves. Practical variational inference for neural networks. NIPS'11, Red Hook, NY, USA, 2011. Curran Associates Inc.

[29] Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cognitive Computation*, 16(1):45–74, 2024.

[30] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating Visual Explanations, 2016.

[31] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations*, 2017.

[32] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network, 2015.

[33] Geoffrey E. Hinton and Drew van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Annual Conference Computational Learning Theory*, 1993.

[34] Sheikh Rabiul Islam, William Eberle, Sheikh Khaled Ghafoor, and Mohiuddin Ahmed. Explainable Artificial Intelligence Approaches: A Survey, 2021.

[35] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding, 2016.

[36] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?, 2017.

[37] Valery Kharitonov, Dmitry Molchanov, and Dmitry Vetrov. Variational Dropout via Empirical Bayes, 2018.

[38] Diederik P. Kingma, Danilo J. Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models, 2014.

[39] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes, 2022.

[40] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial Machine Learning at Scale, 2017.

[41] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles, 2017.

[42] Zachary C. Lipton. The Mythos of Model Interpretability, 2017.

[43] Christos Louizos, Max Welling, and Diederik P. Kingma. Learning Sparse Neural Networks through L_0 Regularization, 2018.

[44] Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions, 2017.

[45] David J. C. MacKay. A practical Bayesian framework for backpropagation networks. *Neural Comput.*, 4(3):448–472, may 1992.

[46] Wesley Maddox, Timur Garipov, Pavel Izmailov, Dmitry Vetrov, and Andrew Gordon Wilson. A Simple Baseline for Bayesian Uncertainty in Deep Learning, 2019.

[47] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks, 2019.

[48] Andrey Malinin and Mark Gales. Predictive Uncertainty Estimation via Prior Networks, 2018.

[49] Brian Mccrindle, Katherine Zukotynski, Thomas Doyle, and Michael Noseworthy. A Radiology-focused Review of Predictive Uncertainty for AI Interpretability in Computer-assisted Segmentation. *Radiology: Artificial Intelligence*, 3:e210031, 09 2021.

[50] Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. Variational Dropout Sparsifies Deep Neural Networks, 2017.

[51] Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022.

[52] Jishnu Mukhoti and Yarin Gal. Evaluating Bayesian Deep Learning Methods for Semantic Segmentation, 2019.

[53] Radford M. Neal. Bayesian Learning for Neural Networks. 1995.

[54] Kazuki Osawa, Siddharth Swaroop, Anirudh Jain, Runa Eschenhagen, Richard E. Turner, Rio Yokota, and Mohammad Emtiyaz Khan. Practical Deep Learning with Bayesian Principles, 2019.

[55] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift, 2019.

[56] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, 2021.

[57] Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. Adversarial Robustness through Local Linearization, 2019.

[58] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models, 2014.

[59] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier, 2016.

[60] Cynthia Rudin. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, 2019.

[61] Waddah Saeed and Christian Omlin. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273, 2023.

[62] Nipuna Sankalpa. Unlocking the Black Box: Explainable Artificial Intelligence (XAI) for Trust and Transparency in AI Systems. *Journal of Digital Art Humanities*, 4:31–36, 06 2023.

[63] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. Adversarial Training for Free!, 2019.

[64] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. 15(1), 2014.

[65] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. ICML'13, pages III–1139–III–1147. JMLR.org, 2013.

[66] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014.

[67] Will Taylor-Melanson, Zahra Sadeghi, and Stan Matwin. Causal Generative Explainers using Counterfactual Inference: A Case Study on the Morpho-MNIST Dataset, 2024.

[68] Ankit Thakkar and Ritika Lohiya. Analyzing fusion of regularization techniques in the deep learning-based intrusion detection system. *International Journal of Intelligent Systems*, 36(12):7340–7388, 2021.

[69] Lin Wang and Kuk-Jin Yoon. Knowledge Distillation and Student-Teacher Learning for Visual Intelligence: A Review and New Outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3048–3068, June 2022.

[70] Florian Wenzel, Jasper Snoek, Dustin Tran, and Rodolphe Jenatton. Hyperparameter Ensembles for Robustness and Uncertainty Quantification, 2021.

[71] Claes Wohlin. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, EASE '14, New York, NY, USA, 2014. Association for Computing Machinery.

[72] Eric Wong and J. Zico Kolter. Provable Defenses Against Adversarial Examples via the Convex Outer Adversarial Polytope, 2018.

[73] Runtian Zhai, Tianle Cai, Di He, Chen Dan, Kun He, John Hopcroft, and Liwei Wang. Adversarially Robust Generalization Just Requires More Unlabeled Data, 2019.

[74] Cheng Zhang, Judith Butepage, Hedvig Kjellstrom, and Stephan Mandt. Advances in Variational Inference, 2018.

A Model Improvement Techniques Categorization

Reference	**Category**	**Relevance**
[24]	Bayesian Approaches and Variational Inference	Probabilistic interpretation aids in understanding and explaining the model's confidence in its predictions.
[8]	Bayesian Approaches and Variational Inference	Incorporates uncertainty in the weights of neural networks, enhancing the model's ability to represent and explain uncertainty.
[39]	Bayesian Approaches and Variational Inference	Method for variational inference in latent variable models, learning disentangled and probabilistically sound representations.
[58]	Bayesian Approaches and Variational Inference	Develops stochastic backpropagation techniques, improving the explainability of generative models through a probabilistic framework.
[26]	Adversarial Robustness	Provides insights into adversarial examples and proposes methods to defend against them, enhancing model robustness and explainability.
[47]	Adversarial Robustness	Proposes adversarial training methods to improve model robustness, making their predictions more reliable and explainable.
[3]	Self-Explaining Models	Proposes self-explaining neural networks, integrating explanation mechanisms into the model architecture.
[32]	Model Compression and Distillation	Introduces knowledge distillation to compress neural networks, making them simpler and more explainable.

Continued on next page

Table 3 – continued from previous page

Reference	Category	Relevance
[30]	Hybrid Techniques	Develops methods for generating visual explanations, combining explainability techniques with visual data.
[16]	Other Methods	Introduces neural ODEs for continuous-time modeling, improving explainability by providing a clear mathematical framework.
[65]	Regularization Techniques	Discusses the role of proper initialization and momentum in training, impacting model stability and interpretability.
[53]	Bayesian Approaches and Variational Inference	Introduces Bayesian methods for neural networks, providing a framework for incorporating uncertainty and enhancing interpretability.
[64]	Regularization Techniques	Proposes dropout as a regularization technique to prevent overfitting, improving model robustness and interpretability.
[28]	Bayesian Approaches and Variational Inference	Presents variational inference methods for neural networks, enhancing model uncertainty quantification and explainability.
[33]	Regularization Techniques	Introduces complexity penalization to promote simpler, more interpretable models.
[66]	Adversarial Robustness	Explores neural networks' vulnerability to adversarial examples, contributing to the development of more robust and interpretable models.
[40]	Adversarial Robustness	Discusses scaling adversarial training, improving model robustness and reliability.
[12]	Adversarial Robustness	Provides methods for robustness evaluation, ensuring valid explanations under adversarial conditions.
[10]	Model Compression and Distillation	Presents methods for compressing models, simplifying their structure and enhancing interpretability.
[25]	Bayesian Approaches and Variational Inference	Applies dropout with learned rates, enhancing uncertainty estimation and explainability by providing probabilistic interpretations.
[22]	Bayesian Approaches and Variational Inference	Comprehensive framework for uncertainty representation in deep learning, improving model transparency and interpretability.
[21]	Uncertainty and Ensembles	Uses deep ensembles to improve predictive uncertainty, enhancing the robustness and interpretability of model predictions.

Continued on next page

Table 3 – continued from previous page

Reference	Category	Relevance
[23]	Bayesian Approaches and Variational Inference	Provides a framework for capturing uncertainty in convolutional networks, improving confidence in model predictions.
[31]	Bayesian Approaches and Variational Inference	Introduces regularization for learning interpretable latent variables, enhancing the modelâs explainability.
[7]	Bayesian Approaches and Variational Inference	Offers a thorough review of variational inference methods, essential for approximating and interpreting complex models.
[38]	Hybrid Techniques	Combines generative models with semi-supervised learning, improving the interpretability of data representations.
[63]	Adversarial Robustness	Presents computationally efficient adversarial training methods, enhancing robustness and the reliability of explanations.
[57]	Adversarial Robustness	Enhances robustness by local linearization, ensuring valid explanations under adversarial conditions.
[4]	Adversarial Robustness	Improves adversarial training methods, addressing overfitting and enhancing the reliability of model explanations.
[73]	Adversarial Robustness	Demonstrates the importance of unlabeled data for robustness, supporting reliable and interpretable model explanations.
[72]	Adversarial Robustness	Introduces a method for provable defenses against adversarial attacks, ensuring stable and reliable model predictions, enhancing explainability through robustness.
[19]	Other Methods	Enhances neural ODEs by adding dimensions to the latent space, improving the model's capacity to capture complex dynamics, and providing a clear mathematical framework.
[9]	Bayesian Approaches and Variational Inference	Proposes a learnable dropout rate within a Bayesian framework, enhancing uncertainty estimation and making model predictions more interpretable and reliable.
[45]	Bayesian Approaches and Variational Inference	Provides a foundational Bayesian framework for neural networks, allowing probabilistic interpretation of weights and outputs, thus improving model transparency.
[17]	Hybrid Techniques	Introduces a method for learning disentangled representations in GANs, enhancing the interpretability of the model's internal representations.

Continued on next page

Table 3 – continued from previous page

Reference	Category	Relevance
[41]	Uncertainty and Ensembles	Uses deep ensembles to estimate predictive uncertainty, improving the robustness and interpretability of model predictions.
[46]	Bayesian Approaches and Variational Inference	Presents a straightforward method for incorporating Bayesian uncertainty, enhancing the interpretability and trustworthiness of the model's predictions.
[43]	Regularization Techniques	Introduces L0 regularization to induce sparsity, simplifying the model and making its decision-making process more understandable.
[36]	Bayesian Approaches and Variational Inference	Distinguishes between different types of uncertainties, providing a framework for better understanding and explaining model predictions, especially in computer vision.
[48]	Uncertainty and Ensembles	Introduces Prior Networks to estimate predictive uncertainty, enhancing model explainability by providing clear confidence levels in predictions.
[70]	Uncertainty and Ensembles	Evaluates multiple hyperparameter configurations to enhance robustness and quantify uncertainty, improving interpretability through robust measures.
[11]	Other Methods	Investigates β-VAE's disentangling properties, enhancing interpretability by aligning latent space with human-interpretable features.
[15]	Other Methods	Explores disentanglement mechanisms in VAEs, improving model explainability by clarifying latent variable roles in output generation.
[74]	Bayesian Approaches and Variational Inference	Reviews advancements in variational inference, enhancing explainability through improved probabilistic approximations and clearer model behavior insights.
[50]	Regularization Techniques	Introduces variational dropout for sparsifying networks, simplifying models and making decision-making processes more transparent.
[55]	Uncertainty and Ensembles	Assesses predictive uncertainty under dataset shifts, ensuring reliability and informativeness of uncertainty estimates for better model explainability.
[54]	Bayesian Approaches and Variational Inference	Integrates Bayesian principles into deep learning, providing probabilistic interpretations and uncertainty estimates to enhance model transparency.

Continued on next page

Table 3 – continued from previous page

Reference	Category	Relevance
[52]	Bayesian Approaches and Variational Inference	Evaluates Bayesian methods for semantic segmentation, providing insights into uncertainty and interpretability improvements for segmentation models.
[35]	Bayesian Approaches and Variational Inference	Integrates Bayesian methods into SegNet, providing uncertainty estimates for scene understanding tasks, enhancing interpretability of model outputs.
[37]	Regularization Techniques	Proposes empirical Bayes approach to variational dropout, sparsifying networks and enhancing explainability by reducing model complexity.

B Impact on Explainability

The table below presents the scores assigned to various papers based on their impact on predictive uncertainty, adversarial robustness, and generative capacity. The scores range from 1 to 5, with 1 indicating minimal impact and 5 indicating significant impact.

Predictive Uncertainty:

- **1:** The technique provides minimal or no insights into the uncertainty of model predictions.
- **2:** The technique offers limited insights into prediction confidence but lacks robust probabilistic interpretation.
- **3:** The technique provides moderate insights into prediction uncertainty with some probabilistic interpretation.
- **4:** The technique offers substantial insights into prediction uncertainty with robust probabilistic interpretation.
- **5:** The technique provides comprehensive probabilistic interpretations, significantly enhancing understanding of model prediction confidence.

Adversarial Robustness:

- **1:** The technique provides minimal or no defense against adversarial attacks.
- **2:** The technique offers limited robustness against adversarial examples but is not comprehensive.
- **3:** The technique provides moderate robustness, improving stability under some adversarial conditions.
- **4:** The technique offers substantial robustness, ensuring stable predictions under various adversarial conditions.
- **5:** The technique provides comprehensive defense mechanisms, significantly enhancing stability and reliability under adversarial conditions.

Generative Capacity:

1: The technique provides minimal or no generative capabilities.

2: The technique offers limited generative capacity.

3: The technique provides moderate generative capacity.

4: The technique offers substantial generative capacity, generating data closely matching the original distribution.

5: The technique provides comprehensive generative capabilities, producing highly realistic and plausible data, significantly enhancing model interpretability.

Reference	**Predictive Uncertainty**	**Adversarial Robustness**	**Generative Capacity**
[24]	5	1	2
[8]	5	2	2
[39]	4	1	5
[58]	4	1	5
[26]	2	5	1
[47]	2	5	1
[3]	3	2	2
[30]	2	2	3
[32]	3	2	2
[16]	3	2	4
[65]	2	3	1
[53]	5	2	3
[64]	5	2	2
[28]	4	2	4
[33]	3	2	2
[66]	2	5	2
[40]	1	5	1
[12]	1	5	1
[10]	2	1	3
[25]	5	1	3
[22]	5	1	2
[21]	5	2	3
[23]	5	2	3
[31]	3	1	5
[7]	4	1	4
[38]	3	1	5
[63]	1	5	1
[57]	1	5	1
[4]	1	5	1
[73]	1	5	1
[72]	1	5	1
[19]	3	1	4
[9]	5	1	3
[45]	5	1	3

Continued on next page

Table 4 – continued from previous page

Reference	Predictive Uncertainty	Adversarial Robustness	Generative Capacity
[17]	1	1	5
[41]	5	2	1
[46]	5	2	1
[43]	3	1	1
[36]	5	1	1
[48]	5	2	1
[70]	5	4	1
[11]	2	1	5
[15]	2	1	5
[74]	4	1	4
[50]	4	2	1
[55]	5	1	1
[54]	5	1	1
[52]	5	1	1
[35]	5	1	1
[37]	4	1	1