

Chapter- 8

Explainable and Interpretable Robust Machine Learning

Thanga Jency T,

*Assistant Professor, Department of Computer Science,
Alpha Arts and Science College,
Chennai, Tamilnadu, India.*

Abstract: The increasing deployment of machine learning systems in high-stakes domains has intensified the need for models that are not only accurate but also transparent and reliable. This chapter explores the integration of explainability, interpretability, and robustness as fundamental pillars of trustworthy artificial intelligence. It begins by introducing core concepts of explainable AI (XAI) and model interpretability, highlighting their distinctions and complementary roles in enhancing human understanding of model behavior. The discussion then extends to robustness, emphasizing its importance in ensuring consistent performance under adversarial perturbations, noisy data, and distributional shifts. The chapter systematically examines interpretability techniques, including intrinsic and post-hoc methods, feature attribution approaches, and visualization strategies, alongside robustness-enhancing techniques such as adversarial training and distributionally robust optimization. It further investigates the interplay between interpretability and robustness, demonstrating how explainability can reveal model vulnerabilities and support the development of more resilient systems. Practical design strategies are presented, including interpretable architectures, regularization methods, and human-in-the-loop approaches, supported by evaluation frameworks that jointly assess transparency and robustness. Applications across critical domains such as healthcare, finance, autonomous systems, and cybersecurity are discussed to illustrate real-world relevance. The chapter also addresses ethical considerations, fairness, and regulatory requirements, underscoring the societal implications of deploying AI systems. Finally, it identifies open research challenges and future directions, emphasizing the need for interdisciplinary collaboration. Overall, this chapter provides a comprehensive foundation for understanding and developing machine learning systems that are explainable, robust, and aligned with the principles of trustworthy AI.

Keywords: *Explainable Artificial Intelligence (XAI), Model Interpretability, Robust Machine Learning, Adversarial Robustness, Distributional Robustness, Feature Attribution, LIME, SHAP, Gradient-Based Explanations, Counterfactual Explanations*

I. INTRODUCTION

The rapid adoption of machine learning (ML) systems across diverse sectors has transformed how decisions are made, predictions are generated, and complex patterns are understood. From diagnosing diseases to automating financial transactions and enabling autonomous navigation, ML models are increasingly embedded in high-stakes environments. However, this widespread deployment has raised two critical concerns: the lack of transparency in model decision-making **and** the vulnerability of models to unreliable or adversarial conditions. These concerns have

given rise to two interconnected research domains—Explainable Artificial Intelligence (XAI) and robust machine learning—both of which are essential for building trustworthy and reliable AI systems.

Explainable Artificial Intelligence (XAI) refers to a set of methods and techniques that make the outputs and inner workings of machine learning models understandable to humans. The goal of XAI is not only to provide predictions but also to offer clear, meaningful justifications for those predictions. This is particularly important for complex models such as deep neural networks, which often operate as “black boxes” with limited transparency. Interpretable machine learning, on the other hand, focuses on designing models that are inherently understandable. These models, such as linear regression, decision trees, and rule-based systems, allow users to directly comprehend how inputs are mapped to outputs without requiring additional explanation tools. Interpretability is often associated with simplicity and transparency, whereas XAI encompasses both inherently interpretable models and post-hoc explanation techniques applied to complex systems. Together, XAI and interpretable machine learning aim to bridge the gap between model performance and human understanding, ensuring that AI systems can be trusted, validated, and effectively governed.

Distinction Between Interpretability and Explainability

Although often used interchangeably, interpretability and explainability represent distinct yet complementary concepts.

- Interpretability refers to the degree to which a human can directly understand the internal mechanics of a model. It is typically associated with models that are simple, transparent, and structured in a way that reveals their decision logic inherently.
- Explainability, in contrast, refers to the ability to provide external explanations for a model’s behavior, particularly when the model itself is too complex to be directly understood. Explainability often relies on post-hoc techniques such as feature attribution, local approximations, or visualization methods.

In essence, interpretability is a property of the model itself, while explainability is a property of the methods used to describe the model. Modern machine learning increasingly relies on highly complex architectures, making explainability techniques indispensable for interpreting their decisions. However, the reliance on post-hoc explanations also introduces challenges related to faithfulness, consistency, and reliability, which must be carefully addressed.

Importance of Robustness in Machine Learning Systems

Robustness in machine learning refers to the ability of a model to maintain reliable performance under varying conditions, including noisy data, distributional shifts, and adversarial perturbations. In real-world scenarios, data is rarely clean or static; instead, it is often subject to uncertainty, incompleteness, and intentional manipulation. Robustness is crucial for several reasons:

- **Resilience to Noise and Errors:** Models must perform reliably even when input data contains inaccuracies or missing values.
- **Adversarial Resistance:** ML systems are increasingly targeted by adversarial attacks, where small, imperceptible changes to input data can lead to incorrect predictions.
- **Generalization Across Environments:** Robust models can adapt to changes in data distribution, ensuring consistent performance across different contexts.
- **Operational Stability:** In production environments, models must handle real-world variability without significant degradation in performance.

Without robustness, even highly accurate models can fail catastrophically when exposed to unexpected or manipulated inputs, undermining their practical utility and trustworthiness.

Motivation: Why Combining Robustness and Interpretability Matters

While interpretability and robustness have traditionally been studied as separate objectives, there is a growing recognition that they are deeply interconnected and mutually reinforcing. A model that is interpretable but not robust may provide clear explanations for incorrect or unstable predictions. Conversely, a robust model that lacks interpretability may produce reliable outputs but fail to offer insights into its decision-making process, limiting trust and accountability. The integration of robustness and interpretability is motivated by several key factors:

- **Trust and Accountability:** Stakeholders require both reliable predictions and understandable reasoning to trust AI systems.
- **Error Diagnosis and Debugging:** Interpretability enables practitioners to identify vulnerabilities and improve robustness.
- **Regulatory Compliance:** Many regulatory frameworks emphasize transparency, fairness, and reliability, necessitating both explainability and robustness.
- **Human-AI Collaboration:** Effective collaboration between humans and AI systems depends on clear communication of model behavior and consistent performance.

Importantly, interpretability can serve as a diagnostic tool for robustness. For example, explanations can reveal whether a model relies on spurious correlations or irrelevant features, which may indicate susceptibility to adversarial manipulation or distributional shifts.

Real-World Implications in Critical Domains

The convergence of explainability and robustness is particularly critical in domains where decisions have significant ethical, financial, or safety implications.

- **Healthcare:** In medical diagnosis and treatment recommendation systems, clinicians must understand the reasoning behind predictions to ensure patient safety. Robustness is equally important, as errors due to noisy or incomplete data can have life-threatening consequences.

- **Finance:** Credit scoring, fraud detection, and algorithmic trading systems require transparency to ensure fairness and regulatory compliance. Robustness is essential to prevent exploitation and maintain stability in dynamic financial markets.
- **Autonomous Systems:** Self-driving vehicles and robotic systems must operate reliably in unpredictable environments. Explainability helps engineers understand system behavior, while robustness ensures safe operation under varying conditions, such as weather changes or sensor noise.

In these domains, the absence of either interpretability or robustness can lead to severe consequences, including loss of trust, regulatory penalties, and potential harm to individuals or society.

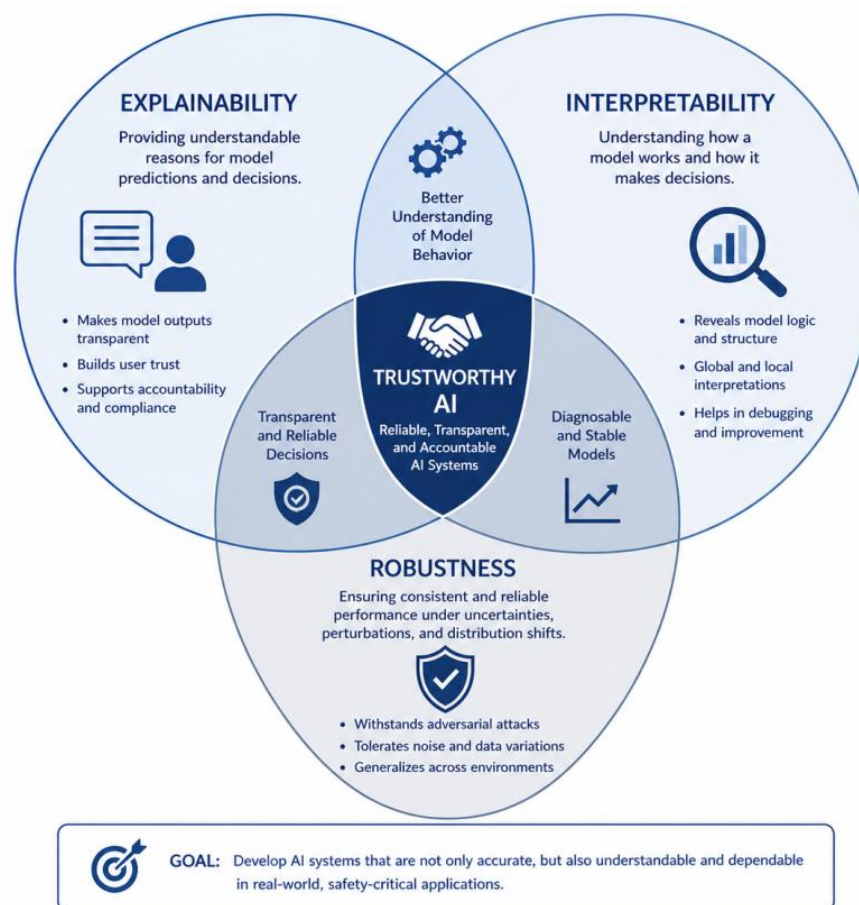


Figure 1: Relationship Between Explainability, Interpretability, and Robustness

II. FUNDAMENTALS OF MODEL INTERPRETABILITY

As machine learning models grow in complexity and capability, understanding how they arrive at their predictions has become a central concern for researchers and practitioners alike. Model interpretability addresses this concern by enabling humans to comprehend, trust, and

effectively manage machine learning systems. It plays a critical role in validating model behavior, diagnosing errors, ensuring fairness, and meeting regulatory requirements. This section explores the foundational concepts of model interpretability, including its types, explanation granularity, feature attribution techniques, and the inherent trade-offs between interpretability and model complexity.

2.1 Types of Interpretability

Model interpretability can broadly be categorized into two major types: intrinsic interpretability and post-hoc interpretability. These approaches differ in how interpretability is achieved – either through model design or through external explanation techniques.

Intrinsic Interpretability

Intrinsic interpretability refers to models that are inherently transparent by design, allowing users to directly understand how inputs are transformed into outputs. These models are typically simpler and structured in a way that exposes their internal logic. Common examples include:

- **Linear Models:** In linear regression and logistic regression, each feature is associated with a coefficient that quantifies its contribution to the prediction. The direction and magnitude of these coefficients provide clear insights into feature influence.
- **Decision Trees:** Decision trees represent decisions as a sequence of hierarchical rules. Each path from the root to a leaf node corresponds to a set of conditions leading to a specific outcome, making the reasoning process easily traceable.
- **Rule-Based Models:** These models use explicit “if-then” rules, offering high transparency and interpretability, especially in domains requiring clear decision logic.

The primary advantage of intrinsic interpretability is its directness and faithfulness, as the explanations are inherently aligned with the model’s structure. However, such models may struggle to capture complex, nonlinear relationships in high-dimensional data, potentially limiting predictive performance.

Post-hoc Interpretability

Post-hoc interpretability involves applying external explanation techniques to understand the behavior of complex, often opaque models such as ensemble methods and deep neural networks. These techniques do not alter the model itself but instead provide approximations or insights into its decision-making process. Popular post-hoc approaches include:

- **Surrogate Models:** Simplified models (e.g., linear models or decision trees) are trained to approximate the predictions of a complex model, enabling easier interpretation.
- **Feature Attribution Methods:** Techniques that quantify the contribution of individual features to a specific prediction.

- **Visualization Tools:** Methods such as saliency maps and activation visualizations that help interpret deep learning models, particularly in image and text domains.

While post-hoc methods enable interpretability for high-performance models, they introduce challenges related to approximation accuracy and faithfulness, as the explanations may not fully reflect the true internal workings of the original model.

2.2 Global vs Local Explanations

Interpretability can also be categorized based on the scope of explanation, namely global and local interpretability.

- **Global Interpretability:** Global explanations aim to describe the overall behavior of a model across the entire dataset. They provide insights into general patterns, feature relationships, and decision boundaries. For example, examining the coefficients of a linear model or the structure of a decision tree offers a global understanding of how the model operates.
- **Local Interpretability:** Local explanations focus on understanding the model's behavior for a specific instance or prediction. These explanations are particularly useful in high-stakes scenarios where individual decisions must be justified. For instance, explaining why a loan application was rejected or why a medical diagnosis was made requires local interpretability.

Both global and local perspectives are essential. Global explanations help in model validation and policy understanding, while local explanations support decision-specific transparency and accountability. In practice, many interpretability frameworks combine both approaches to provide a comprehensive understanding of model behavior.

2.3 Feature Importance and Attribution Methods

A central aspect of model interpretability is identifying which features influence model predictions and to what extent. Feature importance and attribution methods provide quantitative measures of feature contributions, enabling users to understand the driving factors behind model decisions.

Feature Importance

Feature importance measures the relative significance of input variables in determining model predictions. Common approaches include:

- **Model-Specific Importance:** In tree-based models, feature importance can be derived from metrics such as information gain or Gini importance, reflecting how often and effectively a feature is used in decision splits.

- **Permutation Importance:** This model-agnostic method evaluates the decrease in model performance when a feature's values are randomly shuffled. A significant drop in performance indicates high importance.

Feature Attribution Methods

Feature attribution methods provide more granular insights by assigning contribution scores to individual features for specific predictions. Prominent techniques include:

- **SHAP (Shapley Additive Explanations):** Based on cooperative game theory, SHAP values distribute the prediction among features by considering all possible feature combinations, ensuring fairness and consistency.
- **LIME (Local Interpretable Model-agnostic Explanations):** LIME approximates the local behavior of a complex model using a simple, interpretable model, providing insights into individual predictions.
- **Gradient-Based Methods:** In deep learning, gradients are used to estimate how changes in input features affect the output, enabling attribution in high-dimensional spaces.

These methods enhance transparency by revealing which features drive predictions, but they must be carefully interpreted to avoid misleading conclusions, especially when features are correlated or when explanations lack stability.

2.4 Trade-off Between Model Complexity and Interpretability

One of the most fundamental challenges in machine learning is the trade-off between model complexity and interpretability. This trade-off arises because highly complex models, such as deep neural networks and ensemble methods, are capable of capturing intricate patterns and achieving superior predictive performance, but they often lack transparency.

- **Simple Models (High Interpretability):** Models such as linear regression and decision trees are easy to interpret but may suffer from limited expressive power, especially in complex tasks involving nonlinear relationships.
- **Complex Models (High Performance):** Advanced models like deep learning architectures and gradient boosting machines can achieve state-of-the-art performance but operate as black boxes, making their decisions difficult to interpret.

Balancing this trade-off depends on the application context:

- In high-stakes domains (e.g., healthcare, finance), interpretability may take precedence over marginal gains in accuracy.
- In low-risk applications (e.g., recommendation systems), performance may be prioritized, with interpretability addressed through post-hoc methods.

Recent research aims to bridge this gap by developing models that are both accurate and interpretable, as well as improving post-hoc techniques to provide more faithful and reliable

explanations. Additionally, hybrid approaches that combine interpretable components with complex models are gaining traction in both academia and industry.

Model interpretability is a cornerstone of trustworthy machine learning, encompassing both inherently transparent models and techniques for explaining complex systems. By distinguishing between intrinsic and post-hoc interpretability, understanding global and local explanations, and leveraging feature attribution methods, practitioners can gain meaningful insights into model behavior. However, the trade-off between interpretability and complexity remains a persistent challenge, requiring careful consideration based on application requirements. As machine learning continues to evolve, advancing interpretability methods will be essential for ensuring transparency, accountability, and effective human-AI collaboration.

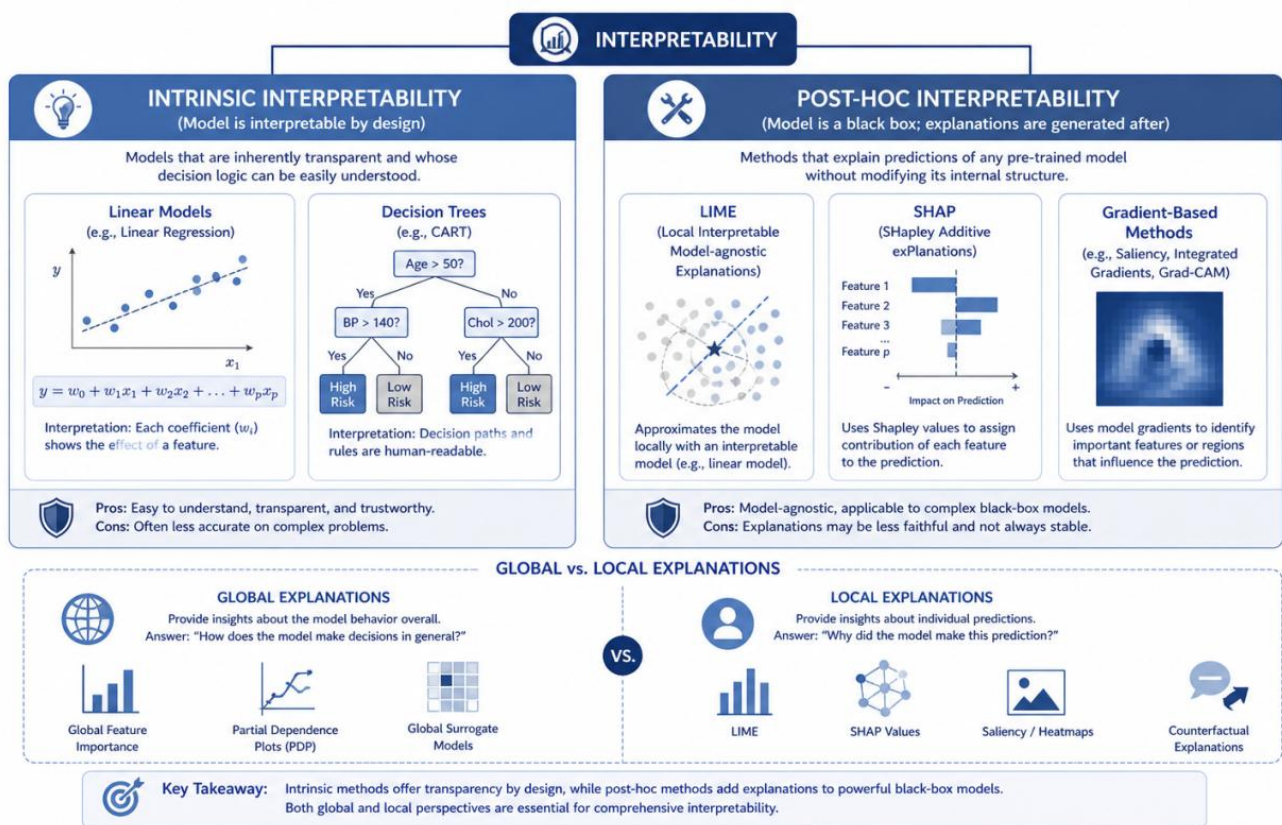


Figure 2: Types of Interpretability (Intrinsic vs Post-hoc)

III. ROBUSTNESS IN MACHINE LEARNING SYSTEMS

As machine learning systems transition from controlled research environments to real-world deployment, their ability to operate reliably under uncertainty becomes a fundamental requirement. Real-world data is often noisy, incomplete, or subject to adversarial manipulation, and deployment conditions frequently differ from those encountered during training. In this context, robustness refers to the capacity of a model to maintain stable and reliable performance despite such variations. This section provides a comprehensive examination of robustness in

machine learning, including its definitions, types, sources of vulnerability, the role of data quality, and commonly used evaluation metrics.

3.1 Definition and Types of Robustness

Robustness in machine learning can be broadly defined as the resilience of a model to changes, uncertainties, and perturbations in input data or environmental conditions. A robust model is expected to generalize well beyond its training distribution and produce consistent outputs even when faced with imperfect or manipulated inputs. Robustness is typically categorized into several types, each addressing a different form of uncertainty or risk.

Adversarial Robustness

Adversarial robustness refers to a model's ability to withstand intentionally crafted perturbations designed to mislead it. Adversarial examples are inputs that have been subtly modified—often imperceptibly to humans—yet cause the model to produce incorrect predictions. For example, in image classification, adding a small, carefully calculated perturbation to an image can lead a deep neural network to misclassify it with high confidence. These vulnerabilities expose critical weaknesses in model decision boundaries and highlight the sensitivity of complex models to small input changes. Ensuring adversarial robustness involves techniques such as adversarial training, gradient regularization, and certified defenses. However, achieving strong adversarial robustness often comes with increased computational cost and potential trade-offs in standard accuracy.

Distributional Robustness

Distributional robustness addresses the challenge of data distribution shifts, where the statistical properties of the input data change between training and deployment. Such shifts can occur due to temporal changes, environmental variations, or differences in data collection processes. Examples include:

- Changes in user behavior over time in recommendation systems
- Variations in medical imaging equipment across hospitals
- Domain shifts between simulated and real-world environments in robotics

A distributionally robust model is designed to perform well across a range of possible data distributions, rather than being optimized solely for a single training distribution. Techniques such as domain adaptation, domain generalization, and robust optimization are commonly employed to address this challenge.

Noise Tolerance

Noise tolerance refers to a model's ability to handle random or unstructured perturbations in data, including measurement errors, missing values, and corrupted inputs. Unlike adversarial

perturbations, noise is typically unintentional and arises from limitations in data collection processes. Examples include:

- Sensor noise in IoT systems
- Labeling errors in annotated datasets
- Transmission errors in communication systems

Robust models should be able to filter out or adapt to such noise without significant degradation in performance. Techniques such as regularization, data augmentation, and robust loss functions are commonly used to improve noise tolerance.

3.2 Sources of Model Vulnerability

Despite advances in machine learning, models remain susceptible to a variety of vulnerabilities that compromise robustness. These vulnerabilities often stem from both data-related and model-related factors.

- **Overfitting to Training Data:** Models that memorize training data rather than learning generalizable patterns are highly sensitive to small perturbations and fail to perform well on unseen data.
- **High Model Complexity:** Complex models, particularly deep neural networks, can learn intricate but fragile decision boundaries that are easily disrupted by small input changes.
- **Spurious Correlations:** Models may rely on irrelevant or coincidental features that do not generalize beyond the training dataset, leading to unreliable predictions under distribution shifts.
- **Insufficient Training Data Diversity:** Lack of representative data can result in models that are poorly equipped to handle variations encountered in real-world scenarios.
- **Weak Regularization and Constraints:** Without appropriate regularization, models may become overly sensitive to noise and adversarial perturbations.
- **Algorithmic Bias and Data Imbalance:** Biases in training data can lead to uneven performance across different subgroups, reducing robustness and fairness.

Understanding these sources of vulnerability is essential for designing strategies that enhance model resilience and reliability.

3.3 Impact of Data Quality and Perturbations

Data quality plays a pivotal role in determining the robustness of machine learning systems. Poor-quality data can significantly degrade model performance and amplify vulnerabilities.

- **Noisy Data:** Random fluctuations in input features can obscure underlying patterns, making it difficult for models to learn meaningful relationships.
- **Missing or Incomplete Data:** Missing values can introduce bias and reduce the reliability of predictions, particularly if not handled appropriately during preprocessing.

- **Corrupted or Inconsistent Data:** Errors in data collection, storage, or transmission can lead to inconsistencies that negatively impact model training and evaluation.
- **Adversarial Perturbations:** Even minor, targeted modifications to input data can cause significant deviations in model output, highlighting the need for robust defenses.
- **Outliers and Anomalies:** Extreme or rare data points can distort model behavior if not properly managed, especially in sensitive applications.

To mitigate these issues, robust machine learning pipelines incorporate data preprocessing, validation, and augmentation techniques. These steps help ensure that models are trained on high-quality, representative datasets and are better equipped to handle real-world variability.

3.4 Metrics for Evaluating Robustness

Evaluating robustness requires specialized metrics that go beyond standard accuracy measures. These metrics assess how well a model performs under different forms of perturbation and uncertainty.

- **Adversarial Accuracy:** Measures model performance on adversarially perturbed inputs, providing insight into resistance against adversarial attacks.
- **Worst-Case Risk:** Evaluates the maximum loss under a set of allowable perturbations, often used in robust optimization frameworks.
- **Robustness Radius:** Quantifies the maximum perturbation magnitude within which the model's prediction remains unchanged.
- **Distributional Generalization Metrics:** Assess performance across different data distributions, such as test sets drawn from shifted domains.
- **Noise Sensitivity Metrics:** Measure performance degradation under varying levels of random noise.
- **Calibration Metrics:** Evaluate the reliability of model confidence scores, ensuring that predicted probabilities reflect true likelihoods even under uncertainty.

A comprehensive evaluation of robustness typically involves stress-testing models under multiple scenarios, including adversarial attacks, noisy inputs, and distributional shifts. This multi-faceted approach provides a more realistic assessment of model performance in real-world conditions. Robustness is a critical property of modern machine learning systems, ensuring reliable performance in the face of uncertainty, variability, and adversarial threats. By understanding the different types of robustness—adversarial, distributional, and noise tolerance—practitioners can design models that are resilient to a wide range of challenges. Identifying sources of vulnerability, maintaining high data quality, and employing appropriate evaluation metrics are essential steps in building robust systems. As machine learning continues to be deployed in high-stakes environments, robustness will remain a cornerstone of trustworthy and dependable AI.

IV. EXPLAINABILITY TECHNIQUES FOR ROBUST MODELS

As machine learning systems become increasingly complex, ensuring both robustness and interpretability requires the use of advanced explanation techniques. These techniques provide insights into how models make decisions, enabling practitioners to assess reliability, detect vulnerabilities, and improve model design. In robust machine learning, explainability is not merely a tool for transparency – it is a critical mechanism for diagnosing weaknesses, validating behavior under perturbations, and enhancing trustworthiness. This section explores key explainability techniques, including model-agnostic methods, gradient-based attribution, counterfactual explanations, visualization approaches, and their role in strengthening robust models.

4.1 Model-Agnostic Methods

Model-agnostic explainability techniques are designed to work with any machine learning model, regardless of its internal structure. These methods treat the model as a black box and generate explanations by analyzing input-output relationships. Their flexibility makes them particularly valuable for interpreting complex models such as ensemble methods and deep neural networks.

LIME (Local Interpretable Model-agnostic Explanations)

LIME is a widely used technique for generating **local explanations** of individual predictions. The core idea behind LIME is to approximate the behavior of a complex model in the vicinity of a specific input instance using a simpler, interpretable model (e.g., a linear model). The process involves:

- Perturbing the input data around the instance of interest
- Observing the corresponding predictions from the original model
- Fitting a simple surrogate model to approximate local behavior

LIME provides interpretable feature weights that indicate the contribution of each feature to the prediction. This enables users to understand why a particular decision was made, such as why a loan application was rejected or why a medical diagnosis was predicted. However, LIME has limitations related to stability and consistency, as different perturbations may yield slightly different explanations. Despite this, it remains a powerful tool for local interpretability.

SHAP (Shapley Additive Explanations)

SHAP is a theoretically grounded method based on cooperative game theory, specifically Shapley values. It assigns each feature a contribution value representing its impact on the model's prediction. Key properties of SHAP include:

- **Additivity:** Feature contributions sum to the model's output
- **Consistency:** Features with greater influence receive higher importance scores

- **Fairness:** Contributions are distributed equitably based on all possible feature combinations

SHAP provides both local and global interpretability, allowing users to analyze individual predictions as well as overall feature importance across the dataset. It is widely used in industry due to its strong theoretical foundation and consistent explanations. Nevertheless, SHAP can be computationally expensive, particularly for high-dimensional data and complex models.

4.2 Gradient-Based Attribution Methods

Gradient-based attribution methods are particularly effective for interpreting deep learning models, where traditional interpretability techniques may fall short. These methods leverage the gradients of the model's output with respect to its input features to estimate feature importance. Common approaches include:

- **Saliency Maps:** Highlight regions of input data (e.g., pixels in an image) that most influence the prediction
- **Integrated Gradients:** Accumulate gradients along a path from a baseline input to the actual input, providing more stable and reliable attributions
- **Gradient-weighted Class Activation Mapping (Grad-CAM):** Produces visual explanations by identifying important regions in convolutional neural networks

These methods are particularly useful in domains such as computer vision and natural language processing, where inputs are high-dimensional and complex. However, gradient-based methods can be sensitive to noise and may sometimes produce misleading explanations if not carefully validated.

4.3 Counterfactual Explanations

Counterfactual explanations focus on answering the question: "What minimal change to the input would alter the model's prediction?" Rather than explaining why a prediction was made, counterfactuals explain how it could have been different. For example:

- In a credit scoring model, a counterfactual explanation might indicate that increasing income by a certain amount would result in loan approval.
- In healthcare, it might suggest changes in clinical parameters that would lead to a different diagnosis.

Key advantages of counterfactual explanations include:

- **Actionability:** They provide clear guidance on how to achieve a desired outcome
- **Model-Agnostic Nature:** They can be applied to any model
- **Human-Centric Interpretability:** They align with human reasoning by focusing on "what-if" scenarios

However, generating realistic and feasible counterfactuals can be challenging, particularly when constraints and domain-specific rules must be considered.

4.4 Visualization Techniques

Visualization plays a crucial role in making complex model behavior more accessible and interpretable. By transforming abstract model computations into visual representations, practitioners can gain intuitive insights into how models process information.

Saliency Maps

Saliency maps highlight the most influential regions of an input, such as pixels in an image or words in a sentence. These maps help identify which parts of the input contribute most to the model's prediction.

Feature Maps and Activation Visualizations

Feature maps visualize intermediate representations within neural networks, particularly in convolutional layers. These visualizations reveal how models extract hierarchical features, such as edges, textures, and higher-level patterns.

Dimensionality Reduction Visualizations

Techniques such as t-SNE and PCA can be used to visualize high-dimensional feature spaces, providing insights into clustering, separability, and potential anomalies in the data. While visualization techniques enhance interpretability, they must be used with caution, as visual patterns may sometimes be misinterpreted without proper quantitative validation.

4.5 Role of Explainability in Detecting Model Weaknesses

Explainability techniques play a pivotal role in identifying and addressing model vulnerabilities, making them essential tools for robust machine learning.

- **Detection of Spurious Correlations:** Explanations can reveal whether a model relies on irrelevant or misleading features, which may compromise generalization.
- **Identification of Adversarial Sensitivity:** By analyzing feature attributions, practitioners can detect inputs that disproportionately influence predictions, indicating susceptibility to adversarial attacks.
- **Bias and Fairness Analysis:** Explainability helps uncover biases in model decisions, enabling corrective measures to ensure fairness and equity.
- **Debugging and Model Improvement:** Insights from explanations can guide feature engineering, model selection, and training strategies to enhance robustness.
- **Validation of Model Behavior:** Explanations allow domain experts to verify whether model decisions align with real-world knowledge and expectations.

Importantly, explainability serves as a bridge between model performance and human understanding, enabling continuous monitoring and improvement of machine learning systems in dynamic environments. Explainability techniques are indispensable for interpreting and improving robust machine learning models. Model-agnostic methods such as LIME and SHAP provide flexible tools for understanding complex systems, while gradient-based attribution and visualization techniques offer deeper insights into neural networks. Counterfactual explanations add a human-centric perspective by highlighting actionable changes. Collectively, these methods not only enhance transparency but also play a critical role in detecting vulnerabilities, diagnosing errors, and strengthening model robustness. As machine learning systems continue to evolve, integrating explainability into the development lifecycle will remain essential for building trustworthy, resilient, and accountable AI systems.

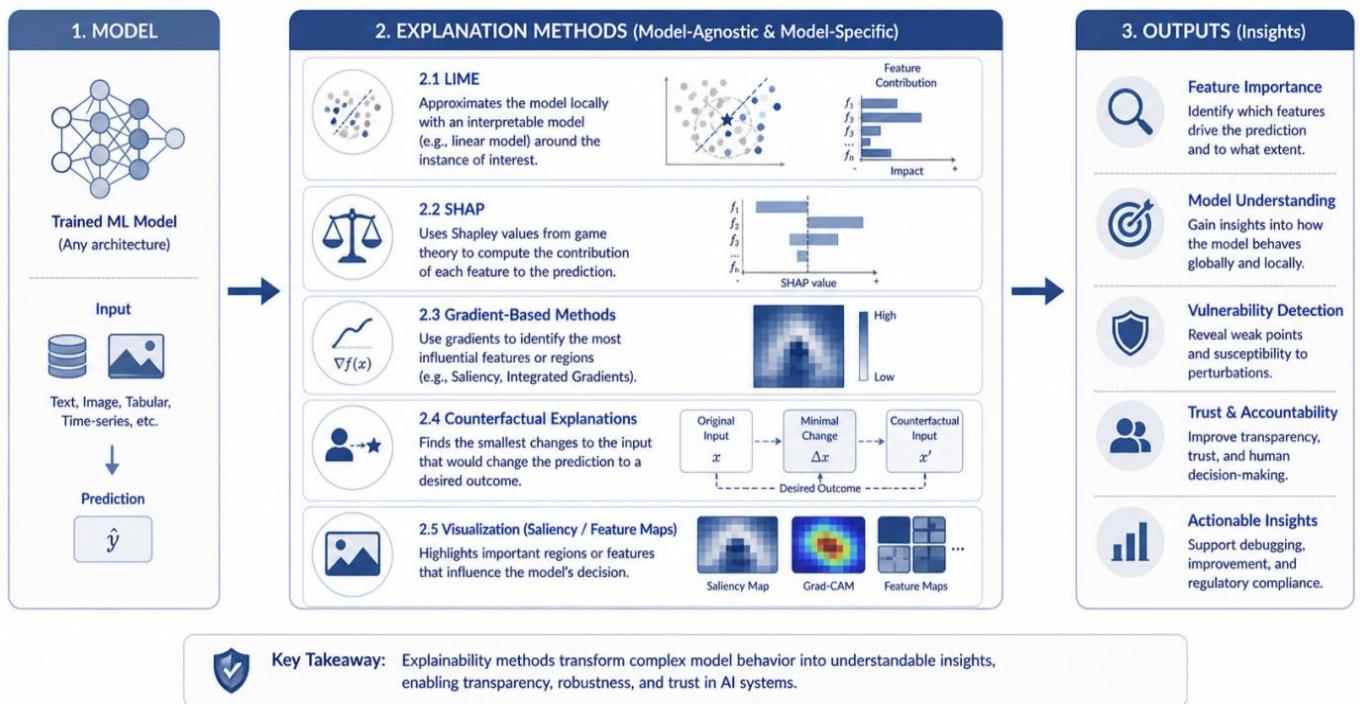


Figure 3: Explainability Techniques Pipeline

V. Interplay Between Interpretability and Robustness

The convergence of interpretability and robustness represents a critical frontier in modern machine learning. While interpretability focuses on making model decisions understandable to humans, robustness ensures that these decisions remain stable and reliable under varying and potentially adversarial conditions. Increasingly, research and industry practice recognize that these two properties are not independent; rather, they are deeply intertwined. Interpretability can reveal weaknesses that compromise robustness, while robust models tend to produce more stable and trustworthy explanations. This section explores their interaction, highlighting how interpretability aids vulnerability detection, the risks posed by adversarial manipulation, the

importance of explanation reliability, and the inherent trade-offs among robustness, accuracy, and interpretability.

5.1 Interpretability as a Tool for Identifying Vulnerabilities

Interpretability provides a powerful lens through which practitioners can examine model behavior and uncover hidden weaknesses. By exposing how features influence predictions, interpretability techniques enable the detection of undesirable patterns, spurious correlations, and over-reliance on irrelevant inputs. For example, in image classification tasks, saliency maps may reveal that a model focuses on background artifacts rather than the object of interest. Similarly, in tabular datasets, feature attribution methods can highlight excessive dependence on proxy variables that may not generalize across datasets. These insights are critical for identifying vulnerabilities that could lead to poor performance under distribution shifts or adversarial conditions.

Moreover, interpretability facilitates model debugging and validation by allowing domain experts to assess whether the model's reasoning aligns with established knowledge. When discrepancies arise, they often signal underlying robustness issues, such as sensitivity to noise or bias in training data. Thus, interpretability acts as an early warning system, enabling proactive improvements to model design and training.

5.2 Adversarial Attacks and Misleading Explanations

The relationship between adversarial robustness and interpretability is complex and, at times, adversarial in itself. Adversarial attacks not only aim to manipulate model predictions but can also distort explanations, leading to misleading interpretations of model behavior. In some cases, adversarial perturbations can cause a model to produce correct predictions while generating incorrect or irrelevant explanations. Conversely, an attacker may craft inputs that preserve the original prediction but significantly alter the explanation, thereby deceiving users about the model's reasoning process. This phenomenon raises serious concerns about the trustworthiness of post-hoc explanations, particularly in high-stakes applications. Furthermore, certain explanation methods are themselves vulnerable to manipulation. For instance, gradient-based attribution techniques can be sensitive to small input changes, making them susceptible to adversarial interference. This dual vulnerability—affecting both predictions and explanations—highlights the need for robust interpretability methods that can withstand adversarial conditions.

5.3 Faithfulness and Reliability of Explanations

A central challenge in interpretability is ensuring that explanations are faithful and reliable representations of the model's true decision-making process.

- Faithfulness refers to the extent to which an explanation accurately reflects the internal workings of the model. An explanation is considered faithful if it correctly identifies the features and interactions that genuinely influence the prediction.

- Reliability to the consistency and stability of explanations across similar inputs or under small perturbations. Reliable explanations should not vary significantly when the input is slightly modified, provided the underlying prediction remains unchanged.

In robust machine learning, faithfulness and reliability are essential for building trust. Unfaithful explanations may provide a false sense of understanding, while unreliable explanations can undermine confidence in the model. To address these challenges, researchers have developed evaluation techniques such as perturbation tests, sensitivity analysis, and consistency checks. Importantly, robustness and interpretability reinforce each other in this context. Robust models tend to produce more stable predictions, which in turn lead to more consistent explanations. Conversely, reliable explanations can help identify and mitigate factors that compromise robustness.

5.4 Trade-offs: Robustness vs Accuracy vs Interpretability

Balancing robustness, accuracy, and interpretability is one of the most significant challenges in machine learning system design. These three objectives often compete, requiring careful consideration of application-specific priorities.

- **Accuracy vs Interpretability:** Highly interpretable models, such as linear models and decision trees, may lack the expressive power needed for complex tasks, resulting in lower predictive accuracy. Conversely, high-performance models like deep neural networks are often less interpretable.
- **Robustness vs Accuracy:** Techniques that improve robustness, such as adversarial training, can sometimes reduce standard accuracy on clean data. This trade-off arises because robust models prioritize stability over fitting to specific patterns in the training data.
- **Robustness vs Interpretability:** Some methods that enhance robustness, such as complex regularization techniques or ensemble methods, may reduce transparency. Conversely, simpler models that are easier to interpret may be less robust to adversarial perturbations or distribution shifts.

In practice, achieving an optimal balance requires context-aware decision-making. For example:

- In healthcare and finance, interpretability and robustness may take precedence over marginal gains in accuracy.
- In large-scale consumer applications, accuracy and scalability may be prioritized, with interpretability addressed through post-hoc methods.

Recent research efforts aim to mitigate these trade-offs by developing models and training strategies that simultaneously enhance robustness and interpretability without significantly compromising accuracy.

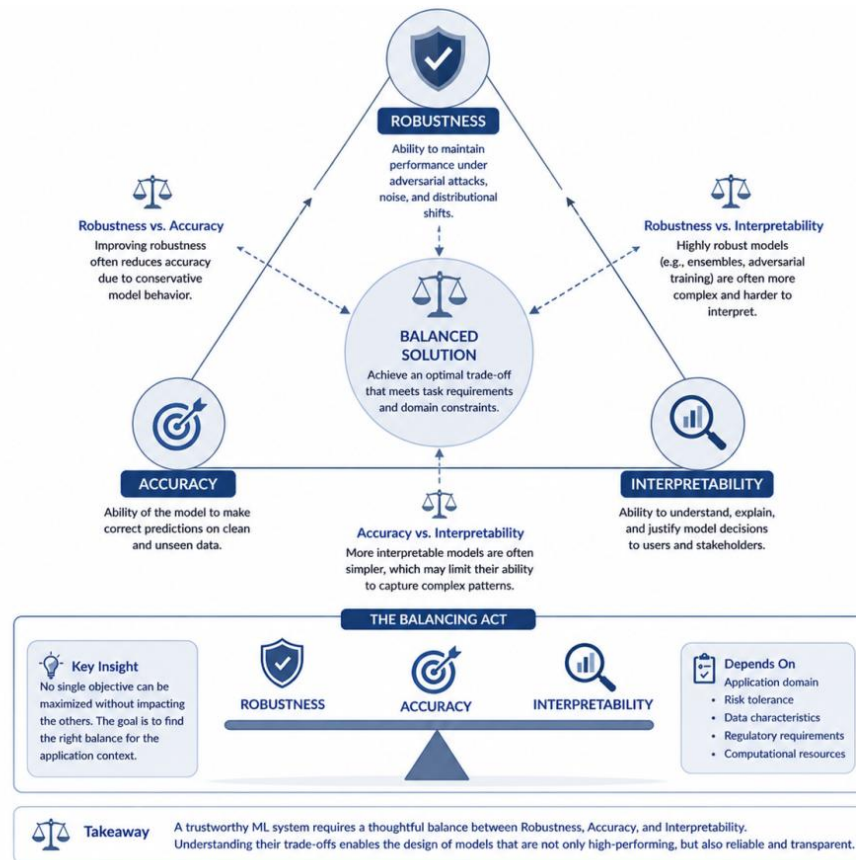


Figure 4: Interplay: Robustness vs Interpretability vs Accuracy

The interplay between interpretability and robustness is a defining characteristic of trustworthy machine learning. Interpretability serves as a diagnostic tool for identifying vulnerabilities, while robustness ensures stable and reliable model behavior under diverse conditions. However, challenges such as adversarial manipulation, explanation reliability, and trade-offs with accuracy necessitate careful consideration in model design and deployment. Through integrated approaches that combine interpretability and robustness, machine learning systems can achieve higher levels of transparency, resilience, and trustworthiness—qualities that are essential for their successful adoption in real-world applications.

5.5 Case Studies Demonstrating the Interaction

The interplay between interpretability and robustness can be observed across various real-world applications, where both properties are critical for reliable performance.

- **Healthcare Diagnostics:** In medical imaging systems, interpretability techniques such as saliency maps have revealed cases where models rely on artifacts (e.g., imaging markers or text annotations) rather than clinically relevant features. Addressing these issues not

only improves interpretability but also enhances robustness to variations in imaging conditions.

- **Financial Risk Assessment:** In credit scoring models, feature attribution methods have uncovered dependencies on sensitive or proxy variables, leading to biased and unstable predictions. By refining feature selection and model design, organizations can improve both fairness and robustness.
- **Autonomous Driving Systems:** Interpretability tools have been used to analyze how self-driving models respond to environmental inputs. In some cases, models were found to misinterpret road signs under adversarial conditions, highlighting vulnerabilities that could compromise safety. Robust training and improved interpretability techniques have been employed to address these challenges.
- **Cybersecurity Applications:** In intrusion detection systems, interpretability helps identify which features contribute to anomaly detection. This insight enables the development of more robust models that can adapt to evolving attack patterns.

These case studies illustrate that interpretability and robustness are mutually reinforcing. By leveraging interpretability to uncover weaknesses and applying robustness techniques to address them, practitioners can build more reliable and trustworthy machine learning systems.

VI. DESIGNING EXPLAINABLE AND ROBUST MODELS

Designing machine learning systems that are both explainable and robust requires a principled integration of modeling choices, training strategies, and evaluation practices. Rather than treating interpretability and robustness as post-hoc add-ons, contemporary approaches embed these properties directly into the model development lifecycle. This section outlines key design paradigms, including robust training with interpretability constraints, interpretable architectures and attention mechanisms, regularization strategies, human-in-the-loop methodologies, and unified evaluation frameworks.

6.1 Robust Training Techniques with Interpretability Constraints

Robust training aims to ensure that model predictions remain stable under perturbations, while interpretability constraints guide models toward transparent and human-understandable decision rules. Integrating these objectives requires careful formulation of training objectives and constraints.

- **Adversarial Training with Explanation Consistency:** Beyond optimizing for correct predictions on adversarially perturbed inputs, models can be trained to maintain **consistent explanations** across clean and perturbed samples. This discourages reliance on brittle or spurious features and promotes stable feature attributions.
- **Right-for-the-Right-Reasons (RRR) Training:** Models are penalized when their explanations (e.g., gradients or attribution maps) highlight features deemed irrelevant by domain knowledge. This aligns model reasoning with expert expectations, improving both interpretability and robustness.

- **Constraint-Based Learning:** Incorporating domain-specific constraints—such as monotonicity (e.g., risk should not decrease as exposure increases) or sparsity (limited number of influential features)—encourages simpler, more interpretable decision boundaries that are less sensitive to noise.
- **Distributionally Robust Optimization (DRO) with Transparency Objectives:** DRO frameworks can be augmented with interpretability-aware penalties, ensuring that the model performs well across distribution shifts while maintaining stable and explainable feature contributions.

These approaches emphasize that robustness is not only about performance under perturbation but also about stability in the reasoning process.

6.2 Incorporating Attention Mechanisms and Interpretable Architectures

Model architecture plays a central role in determining both interpretability and robustness. Designing inherently interpretable structures reduces reliance on post-hoc explanations and enhances transparency.

- **Attention Mechanisms:** Attention layers, widely used in natural language processing and computer vision, assign weights to different input components, indicating their relative importance. When properly calibrated, attention can provide intuitive insights into model focus. However, it is important to note that attention weights do not always equate to causal importance, necessitating careful validation.
- **Self-Explaining Neural Networks (SENN):** These architectures explicitly decompose predictions into interpretable components, such as concept-based representations and relevance scores, ensuring that explanations are built into the model.
- **Generalized Additive Models (GAMs) and Neural Additive Models (NAMs):** These models express predictions as sums of individual feature contributions, combining interpretability with flexibility. Their modular structure enhances robustness by isolating the influence of each feature.
- **Prototype and Case-Based Models:** These models make predictions by comparing inputs to representative examples (prototypes), enabling intuitive explanations and improved robustness through similarity-based reasoning.

6.3 Regularization for Stability and Transparency

Regularization techniques are essential for controlling model complexity and improving generalization. When designed appropriately, they also enhance interpretability and robustness.

- **L1 and Sparsity-Inducing Regularization:** Encourages models to rely on a smaller subset of features, making explanations more concise and reducing sensitivity to irrelevant inputs.

- **Gradient Regularization:** Penalizes large gradients with respect to input features, promoting smoother decision boundaries and reducing susceptibility to adversarial perturbations.
- **Lipschitz Regularization:** Enforces constraints on how much the model output can change in response to input variations, directly improving robustness and stabilizing explanations.
- **Explanation Regularization:** Explicitly constrains the behavior of explanation methods (e.g., SHAP or gradient-based attributions) during training to ensure consistency and alignment with domain knowledge.
- **Dropout and Noise Injection:** Introducing controlled noise during training improves generalization and helps models become less sensitive to input perturbations.

These regularization strategies collectively contribute to stable, interpretable, and generalizable models, which are essential for deployment in dynamic environments.

6.4 Human-in-the-Loop Approaches

Human-in-the-loop (HITL) methodologies integrate human expertise into the model development and deployment process, enhancing both interpretability and robustness.

- **Interactive Model Refinement:** Domain experts review model explanations and provide feedback, which is used to adjust model parameters, feature selection, or training objectives.
- **Annotation and Constraint Specification:** Experts can label not only data but also **feature relevance**, guiding the model toward more meaningful and robust decision patterns.
- **Active Learning with Interpretability Feedback:** Models identify uncertain or ambiguous instances, and human input is used to improve both predictive performance and explanation quality.
- **Monitoring and Auditing:** Continuous human oversight ensures that models behave as expected over time, particularly in the presence of distribution shifts or emerging risks.

HITL approaches are particularly valuable in high-stakes domains, where human judgment is essential for validating model decisions and ensuring accountability.

6.5 Evaluation Frameworks Combining Robustness and Explainability

Traditional evaluation metrics, such as accuracy or precision, are insufficient for assessing explainable and robust models. Comprehensive evaluation frameworks must consider both predictive performance and quality of explanations under varying conditions.

- **Robustness Evaluation:**
 - Performance under adversarial attacks
 - Sensitivity to noise and perturbations
 - Generalization across distribution shifts

- **Explainability Evaluation:**
 - **Faithfulness:** Do explanations accurately reflect model behavior?
 - **Stability:** Are explanations consistent across similar inputs?
 - **Sparsity and Simplicity:** Are explanations concise and interpretable?
- **Joint Metrics:**
 - **Explanation Consistency Under Perturbation:** Measures whether explanations remain stable when inputs are slightly modified
 - **Robust Feature Attribution Scores:** Evaluate how reliably feature importance persists across different scenarios
 - **Human Evaluation Metrics:** Assess interpretability through user studies and expert feedback
- **Benchmarking and Stress Testing:**

Models should be evaluated across multiple datasets, perturbation scenarios, and explanation methods to ensure comprehensive assessment.

An integrated evaluation framework enables practitioners to quantify trade-offs and ensure that improvements in robustness do not come at the expense of interpretability, and vice versa. Designing explainable and robust machine learning models requires a holistic approach that integrates training strategies, architectural choices, regularization techniques, human expertise, and comprehensive evaluation frameworks. By embedding interpretability constraints into robust training, leveraging interpretable architectures, and employing stability-enhancing regularization, practitioners can develop models that are both transparent and resilient. Human-in-the-loop approaches further enhance trust and accountability, while unified evaluation frameworks ensure that both robustness and explainability are rigorously assessed. As machine learning systems continue to evolve, these design principles will play a pivotal role in advancing trustworthy and reliable AI.

VII. APPLICATIONS AND EMERGING RESEARCH DIRECTIONS

The integration of explainability and robustness has transitioned from a theoretical aspiration to a practical necessity across multiple high-impact domains. As machine learning systems increasingly influence critical decisions, stakeholders demand models that are not only accurate but also transparent, reliable, and accountable. This section examines key application areas where explainable robust AI is essential, followed by a discussion of ethical considerations, regulatory frameworks, and emerging research challenges shaping the future of this field.

7.1 Explainable Robust AI in Key Application Domains

In healthcare, machine learning models are widely used for disease diagnosis, medical imaging analysis, and treatment recommendation. The integration of explainability and robustness is crucial due to the life-critical nature of decisions. Explainability enables clinicians to:

- Understand the rationale behind model predictions
- Validate model outputs against medical knowledge
- Identify potential biases or errors in diagnosis

Robustness ensures that models:

- Perform reliably across diverse patient populations
- Handle noisy or incomplete medical data
- Maintain accuracy under varying imaging conditions or sensor noise

For instance, in radiology, explainability techniques such as saliency maps help highlight regions of medical images that influence predictions, while robust training ensures consistent performance across different imaging devices and conditions. Together, these capabilities enhance clinical trust and facilitate adoption in practice.

Financial Risk Assessment

Machine learning plays a central role in financial applications such as credit scoring, fraud detection, and algorithmic trading. In this domain, explainability and robustness are essential for transparency, fairness, and regulatory compliance. Explainability supports:

- Justification of credit decisions to customers and regulators
- Detection of biased or discriminatory patterns
- Improved stakeholder trust in automated decision systems

Robustness ensures:

- Stability under market fluctuations and economic shifts
- Resistance to adversarial manipulation in fraud detection systems
- Consistent performance across diverse customer segments

For example, feature attribution methods can reveal which financial factors contribute to a credit decision, while robust models can adapt to changing economic conditions without significant degradation in performance.

Autonomous Vehicles

Autonomous systems, including self-driving vehicles, operate in highly dynamic and uncertain environments. In such systems, both explainability and robustness are critical for safety and reliability. Explainability enables:

- Engineers to understand system behavior and decision pathways
- Debugging and validation of perception and control modules
- Transparent communication of decisions in safety-critical scenarios

Robustness ensures:

- Reliable operation under varying environmental conditions (e.g., weather, lighting)
- Resistance to sensor noise and adversarial inputs (e.g., manipulated road signs)

- Stable performance in complex and unpredictable traffic scenarios

The combination of these properties is essential for ensuring that autonomous systems can be trusted in real-world deployments.

Cybersecurity

In cybersecurity, machine learning models are used for intrusion detection, malware classification, and anomaly detection. These systems must be both explainable and robust to counter increasingly sophisticated threats. Explainability helps:

- Analysts understand why certain activities are flagged as malicious
- Identify patterns and signatures associated with cyber threats
- Improve incident response and forensic analysis

Robustness ensures:

- Detection of novel and evolving attack patterns
- Resistance to adversarial evasion techniques
- Stability under noisy and incomplete network data

For instance, interpretable models can highlight suspicious features in network traffic, while robust training enables systems to adapt to rapidly changing threat landscapes.

7.2 Ethical Considerations and Fairness

The deployment of explainable and robust AI systems raises important ethical considerations, particularly in relation to fairness, accountability, and transparency.

- **Bias and Discrimination:** Models trained on biased data may produce unfair outcomes. Explainability techniques can help identify such biases, while robustness ensures consistent performance across different demographic groups.
- **Accountability:** Transparent models enable stakeholders to assign responsibility for decisions, which is essential in regulated industries.
- **Trust and User Acceptance:** Users are more likely to trust AI systems that provide clear explanations and demonstrate reliable performance.
- **Privacy Concerns:** Some explanation methods may inadvertently expose sensitive information, necessitating careful design to balance transparency and privacy.

Ethical AI design requires integrating fairness constraints, conducting bias audits, and ensuring that explanations are both meaningful and responsible.

7.3 Regulatory Requirements and Standards

As AI systems become more pervasive, governments and regulatory bodies are establishing frameworks to ensure their safe and ethical use. Explainability and robustness are central to many of these regulations.

- **Transparency Requirements:** Regulations increasingly mandate that automated decisions be explainable, particularly in domains such as finance and healthcare.
- **Fairness and Non-Discrimination:** Models must demonstrate equitable performance across different groups, requiring both interpretability and robustness.
- **Auditability and Accountability:** Organizations must maintain records of model behavior, training data, and decision logic to enable audits and compliance checks.
- **Safety Standards:** In domains such as autonomous systems, rigorous testing and validation are required to ensure robustness under diverse conditions.

7.4 Open Research Challenges and Future Directions

Despite significant progress, several challenges remain in the development of explainable and robust machine learning systems.

- **Scalability of Explanation Methods:** Many interpretability techniques are computationally intensive, limiting their applicability to large-scale systems.
- **Faithfulness of Explanations:** Ensuring that explanations accurately reflect model behavior remains an open challenge, particularly for complex models.
- **Robustness of Explanations:** Explanation methods themselves must be resilient to adversarial manipulation and input perturbations.
- **Unified Frameworks:** Developing integrated approaches that simultaneously optimize for accuracy, robustness, and interpretability is an ongoing research area.
- **Human-Centered Evaluation:** Measuring the effectiveness of explanations from a human perspective requires interdisciplinary research involving psychology and human-computer interaction.
- **Cross-Domain Generalization:** Ensuring that models remain robust and interpretable across different domains and environments is a key challenge for real-world deployment.

VIII. CONCLUSION

The convergence of interpretability, explainability, and robustness represents a foundational paradigm for the development of trustworthy machine learning systems. Throughout this chapter, interpretability has been examined as the capacity of models to offer inherent transparency, while explainability extends this capability by enabling post-hoc insights into complex model behavior. Robustness, in parallel, ensures that machine learning systems maintain consistent and reliable performance under uncertainty, including noisy data, adversarial perturbations, and distributional shifts. Together, these three dimensions form a cohesive framework that addresses both the comprehensibility and resilience of modern AI

systems. Integrating transparency with robustness is not merely a desirable feature but a critical requirement for real-world deployment. Machine learning models increasingly operate in high-stakes environments where errors can have significant ethical, financial, or safety implications. In such contexts, it is insufficient for models to achieve high predictive accuracy alone; they must also provide clear, interpretable reasoning and demonstrate stability under diverse and unpredictable conditions. The alignment of explainability with robustness enables practitioners to detect vulnerabilities, validate model behavior, and ensure that decisions are both justifiable and dependable. This integration also supports regulatory compliance, facilitates human-AI collaboration, and enhances stakeholder trust. Despite significant advancements, the field continues to face important research challenges. Ensuring the faithfulness and stability of explanations, developing scalable and efficient interpretability techniques, and achieving robustness without compromising accuracy remain open problems. Moreover, the inherently interdisciplinary nature of these challenges calls for collaboration across domains, including computer science, statistics, ethics, law, and human-computer interaction. Future progress will depend on the development of unified frameworks that seamlessly integrate interpretability and robustness, as well as the establishment of standardized evaluation protocols and benchmarks.

REFERENCES

- [1]. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations (ICLR).
- [2]. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. International Conference on Learning Representations (ICLR).
- [3]. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations (ICLR).
- [4]. Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. (2016). DeepFool: A simple and accurate method to fool deep neural networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2574–2582.
- [5]. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.
- [6]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS), 4765–4774.
- [7]. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
- [8]. Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61(10), 36–43.
- [9]. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. International Conference on Machine Learning (ICML), 3319–3328.
- [10]. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. IEEE International Conference on Computer Vision (ICCV), 618–626.

- [11]. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2), 841–887.
- [12]. Arjovsky, M., Bottou, L., Gulrajani, I., & Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.
- [13]. Duchi, J. C., Glynn, P. W., & Namkoong, H. (2021). Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3), 946–969.
- [14]. Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *International Conference on Learning Representations (ICLR)*.
- [15]. Ross, A. S., Hughes, M. C., & Doshi-Velez, F. (2017). Right for the right reasons: Training differentiable models by constraining their explanations. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2662–2670.
- [16]. Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.
- [17]. Yeh, C.-K., Kim, J., Yen, I. E., & Ravikumar, P. (2019). On the (in)fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems (NeurIPS)*, 10965–10976.
- [18]. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
- [19]. Molnar, C. (2022). *Interpretable machine learning* (2nd ed.). Lulu.com.
- [20]. Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities*. MIT Press.