On Assessing ML Model Robustness: A Methodological Framework

Afef Awadid[1]
IRT SystemX, Palaiseau, France

Boris Robert
IRT Saint Exupéry, Toulouse, France

Abstract

Due to their uncertainty and vulnerability to adversarial attacks, machine learning (ML) models can lead to severe consequences, including the loss of human life, when embedded in safety-critical systems such as autonomous vehicles. Therefore, it is crucial to assess the empirical robustness of such models before integrating them into these systems. ML model robustness refers to the ability of an ML model to be insensitive to input perturbations and maintain its performance. Against this background, the Confiance.ai research program proposes a methodological framework for assessing the empirical robustness of ML models. The framework encompasses methodological processes (guidelines) captured in Capella models, along with a set of supporting tools. This paper aims to provide an overview of this framework and its application in an industrial setting.

2012 ACM Subject Classification Software and its engineering → Software verification and validation

Keywords and phrases ML model robustness, assessment, framework, methodological processes, tools

Digital Object Identifier 10.4230/OASIcs.SAIA.2024.1

Category Academic Track

Funding This work has been supported by the French government under the "France 2030" program.

Acknowledgements We would like to extend special thanks to the robustness team of the confiance.ai research program, composed of AI engineers, for producing the deliverables that served as the foundation for this paper.

1 Introduction

Central to Machine Learning (ML) is a "data-driven AI technology that automatically discovers patterns and relationships from large volumes of data using algorithmic models" [27]. In this context, ML enables computers to learn complex statistical patterns from data and make decisions without explicit programming [14]. Moreover, it facilitates tackling "tasks that are too difficult to solve with traditional programming paradigms" [17]. ML has also proven effective in analyzing customer demand, allowing for accurate anticipation and planning of future needs [16]. It is therefore unsurprising that ML has significantly transformed various industries [11]. However, the inherent risks and uncertainties associated with ML technology pose significant challenges, especially when implementing it in safety-critical systems such as autonomous vehicles.

Indeed, the successful deployment of ML models can be constrained by biases present in the training data. If the data used for training is not representative of the actual scenarios that the system will encounter in the real world, the ML model may make inaccurate and erroneous decisions. Such inaccuracies can lead to catastrophic outcomes, such as vehicle

[1] corresponding author

accidents, posing a significant threat to human life [4]. Furthermore, the complexity of ML-based safety-critical systems makes them susceptible to various adversarial attacks throughout the ML pipeline. This vulnerability stems from the unpredictable and dynamic environments in which these systems operate, where even minor changes in input data can result in serious consequences [24].

To address this challenge, the Confiance.ai research program has placed particular emphasis on the robustness of ML models by developing a framework to support their assessment. This framework includes methodological guidelines/processes captured in Capella models, as well as a set of tools. It is part of a comprehensive methodology designed to guide the development of trustworthy ML-based systems.

This paper aims to provide an overview of the proposed framework. It, thus seeks to address the following research question: How can we support the assessment of ML models' robustness to ensure their successful deployment in ML-based safety-critical systems?

The rest of the paper is structured as follows. Section 2 introduces the theoretical background of this work. The methodological framework for assessing ML model robustness is presented in Section 3. Section 4 concludes the paper with suggestions for future work.

2 Theoretical Background

2.1 Context and Motivation

Integrating ML techniques into safety-critical systems can promote autonomy and support decision-making processes. However, this increased autonomy may also introduce unpredictability and uncertainty, which could challenge the system's reliability. Accordingly, to ensure the trustworthiness of ML-based safety-critical systems, it is essential to assess the robustness of ML models before deploying them operationally.

Trustworthiness refers to the probability that a system has certain established properties such as robustness and reliability, with higher trustworthiness indicating a greater likelihood that the system will exhibit those properties [5].

Against this backdrop, the Confiance.ai research program has proposed an end-to-end method aimed at guiding the development of ML-based safety-critical systems, with the aim of bolstering the French industry's confidence in these technologies. This method encompasses methodological processes and associated tools, covering the entire lifecycle of ML-based systems, from design to operation. It addresses critical aspects of ML model trustworthiness, such as robustness and explainability. This paper specifically focuses on the part of the end-to-end method dedicated to the evaluation of ML model robustness, highlighting the methodological framework developed for this purpose.

2.2 Related Work

Robustness refers to the ability of an AI system to maintain its level of performance under any circumstances (e.g. external interference or harsh environmental conditions) [2]. It has emerged as a key quality characteristic of trustworthy AI. In this context, [3] emphasizes the importance of robustness as a sub-characteristic of reliability in the quality of AI systems. One reason for this emphasis is that robustness allows AI systems to maintain normal operation and avoid unreasonable safety risks throughout their lifecycle, even in the case of misuse and other unfavorable conditions [1].

Given this, it is not surprising that the robustness assessment of ML models has been attracting significant attention, leading to several approaches in this regard. These approaches can be broadly categorized into two main types: 1) formal verification-based approaches and

2) statistical verification-based approaches. In the first category, robustness verification is formulated as a satisfiability problem, where the goal is to identify the smallest set of inputs that satisfy the robustness condition and result in a bound for local robustness (i.e., the model's ability to maintain its output within specific regions in the input space). Examples of these approaches include stability region verification [9, 7], sensitivity analysis [20], and safety region verification [12, 21, 26].

The second category of approaches relies on statistical verification to assess local robustness. These methods quantify the local robustness of ML models by evaluating the probability of inputs violating or satisfying the robustness condition within the verification region. Such approaches are generally implemented based on input domain sampling, as seen in [8, 13], or within a formal framework (e.g., [23, 25, 6]).

In view of this, existing approaches for ML model robustness assessment tend to address this issue from a purely technical point of view. Consequently, the proposed solutions are tailored to individuals with high technical skills, specifically ML algorithm engineers. Implementing these solutions requires substantial knowledge in ML algorithm engineering and related fields. In contrast, within the Confiance.ai research program, we provide both methodological support (engineering processes/guidelines) and technical support (tools for those guidelines) to assess ML model robustness. Our resulting methodological framework is part of an end-to-end method that covers the entire ML systems engineering lifecycle and aims to guide the development of trustworthy ML systems. Therefore, it is designed to be accessible to various engineering specialists, such as system engineers, ML algorithm engineers, and data engineers.

2.3 An Overview of the End-to-End Methodology

Within the Confiance.ai research program, an end-to-end methodology has been developed by a large and diverse group of experts including industry actors to provide a set of methodological processes/guidelines aimed at assisting in the development of reliable and trustworthy ML-based safety-critical systems. These processes seek to cover the entire lifecycle of ML-based systems (see Figure 1).

Figure 1 outlines the main engineering steps of our methodology, which are defined in line with the ISO/IEC 5338:2023 standard. The latter describes the life cycle of AI systems based on machine learning and heuristic systems. It illustrates the lifecycle of systems engineering that integrates Machine Learning (ML) into the traditional V cycle of system development, creating a customized W cycle for software engineering processes. This W cycle highlights the critical step of evaluating the ML model for reliability at the algorithm level before its implementation at the software level. In this paper, we focus on the engineering activity "Evaluation of ML Model and its Complementary Software Items" in Figure 1.

3 A Methodological Framework for Assessing ML Model Robustness

3.1 Engineering Processes

To evaluate the robustness of ML models, the Confiance.ai research program proposes two engineering/methodological processes. These processes have been captured in Capella models and were developed collaboratively by systems engineers, ML algorithm engineers, and data engineers from various academic and industrial partners of the Confiance.ai research program. The Capella tool was chosen due to its widespread use and familiarity among our multidisciplinary team.

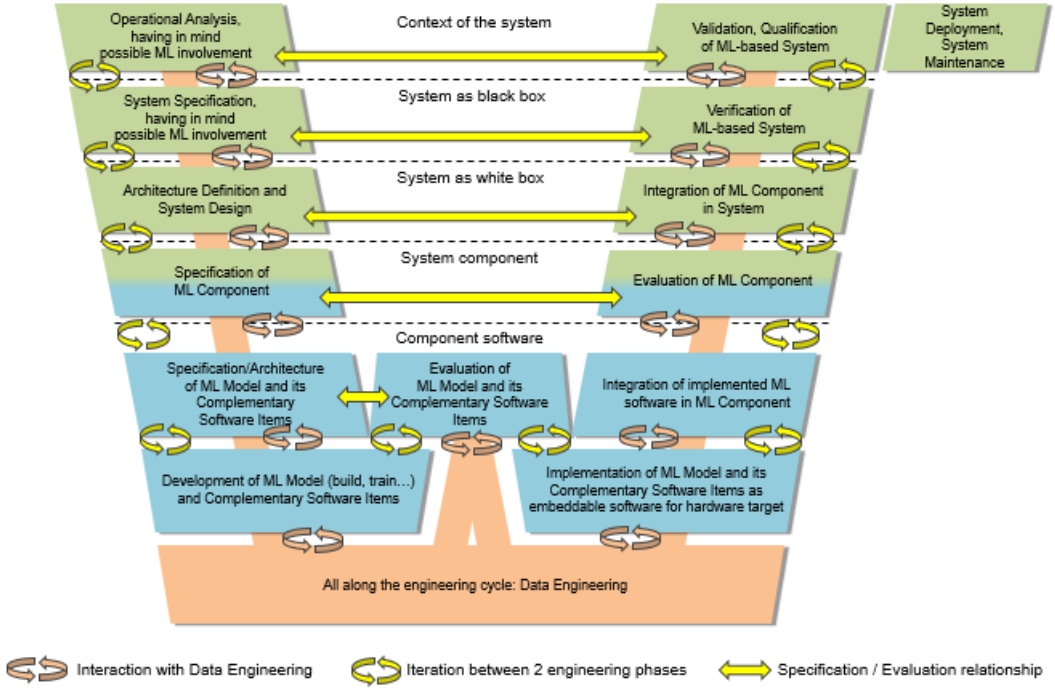

Figure 1 Lifecycle of ML-based systems.

The proposed engineering processes consider two strategies for assessing ML model robustness: 1) robustness testing by sampling and perturbation (also known as empirical robustness) and 2) robustness through formal evaluation (also known as formal robustness). The former involves constructing test datasets with perturbations to evaluate the correctness of a model's output in response to these perturbed inputs. The latter aims to determine if a model is robust within a specific perturbation range and has made significant progress through formal and statistical verification techniques [22].

Both strategies can be combined to assess the robustness of an ML model. However, it is important to note that the formal robustness strategy may not be feasible for certain types of models due to their complexity or lack of formal specifications. In this sense, the adoption of formal verification to evaluate the robustness of ML models depends on certain constraints such as the acceptability of formal proofs, the compatibility of verification tools with the ML model algorithm, and the dimension of the data space. For this reason, the scope of this paper will focus on the robustness testing by sampling and perturbation strategy.

The engineering/methodological process capturing the empirical robustness strategy (i.e., robustness by sampling and perturbation) is presented in Figure 2. A prerequisite of this strategy is that the ML model robustness requirement is expressed as a maximum tolerable deviation of the ML model's behavior in response to a certain intensity of perturbation in the input data.

As shown in Figure 1, the process of evaluating ML model robustness is conducted by the ML algorithm engineer through sampling and perturbation techniques. The engineering activity, titled "Evaluate the robustness of the trained ML model using sampling and perturbation tests," is a sub-activity of the broader task "Evaluate the trained ML model (and its complementary software elements, if necessary)." This indicates that robustness evaluation via sampling and perturbation represents one of several strategies for assessing the robustness of a trained ML model.

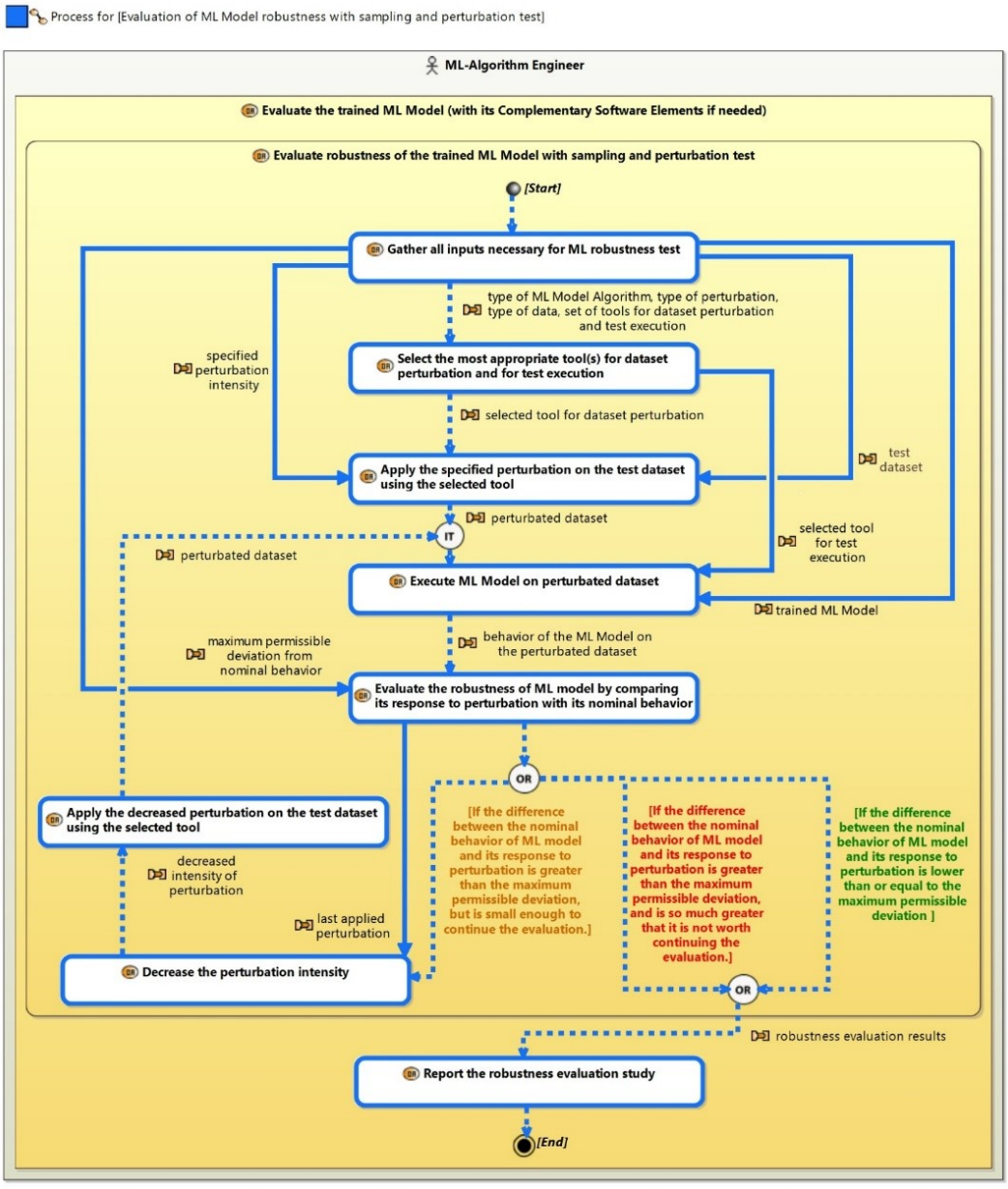

Figure 2 Process for evaluating ML model robustness by sampling and perturbation.

Referring to Figure 2, the process of evaluating ML model robustness according to the empirical robustness strategy includes the following steps:

- Gather all necessary inputs for the ML model robustness test: This initial activity involves collecting all essential information required for testing the robustness of the ML model such as the problem type (e.g. image classification), the data type (e.g. image), the type of perturbation (e.g. variations in image luminosity).
- The second step involves selecting the most appropriate tool for dataset perturbation and test execution from the Confiance.ai set of tools based on the relevant inputs.
- Using the selected tool, the specified perturbation is applied to the test dataset. The tool then executes the test and measures the ML model's performance against predefined KPIs. The resulting behavior of the ML Model is thus captured.
- The robustness of the ML Model is evaluated by comparing this perturbed response with its nominal behavior on the non-perturbed Test Dataset. This involves looking for deviations from expected behavior within permissible limits.
- If the difference between the perturbed response of the ML Model and its nominal behavior is lower than or equal to the given maximum permissible threshold, the robustness evaluation study should report that the initial perturbation level given as input has permitted to reach the target level of robustness. This means that the ML Model robustness requirement is satisfied.
- If such difference is slightly greater than the given maximum permissible threshold, the intensity of the perturbation on the test dataset can be decreased until the difference between the nominal behavior of the ML Model and its response to disturbance is smaller than the specified tolerance value. An alternative approach is to continue the robustness evaluation of the ML model until the perturbation level approaches zero, irrespective of the threshold set for the impact. This method provides a more comprehensive analysis of the model's robustness range.
- If the difference is significantly greater than the maximum permissible deviation, the evaluation study can be reported, as it is not worthwhile to continue the evaluation.

3.2 Supporting Tools

To facilitate the application of the proposed methodological process for evaluating ML model robustness in an industrial setting, the Confiance.ai research program provides a set of supporting tools and components. These components are designed to assess the robustness of trained models against various perturbations. These perturbations can be specific to images, such as Gaussian blur or geometric transformations, and can also include adversarial attacks.

3.2.1 Component 331: Adversarial Attack Characterization Component

This component aims to evaluate the reproducibility level of a decision model by providing its robustness ratio based on specified interest variables, including the variation of their intensity. It relies on the open-source ART-IBM library [19] and evaluates a neural network model against a set of adversarial attacks, such as Projected Gradient Descent (PGD) [15], DeepFool [18], and NewtonFool [10]. The Adversarial Attack Characterization Component has been successfully applied to various use cases.

3.2.2 Component 332: AI Metamorphis Observer Component (AIMOS)

The AIMOS component evaluates metamorphic properties in AI models. This toolkit is designed to be agnostic, enabling comparisons across a wide range of model types and use cases. It allows testing various metamorphic properties and transformations over different

value ranges and models, with some visual displays included. AIMOS features several types of attacks, such as Gaussian noise (electrically-induced noise), Poisson noise (thermally-induced noise), Gaussian blur (camera vibration), vertical and horizontal motion blur (camera vibration), pixel, column, and line loss (camera sensor failure), and defocus blur (camera focus variation).

3.2.3 Component 333: Amplification Method for Robustness Evaluation Component

The Amplification Method for Robustness Evaluation Component assesses the robustness of models using image and time series data by applying amplification methods with noise functions to the dataset. These noise functions are implemented as Python scripts. For image datasets, the provided noise functions include Gaussian blur, horizontal and vertical motion blur, dead pixels, lines and columns, additive Gaussian noise, and multiplicative Poisson noise.

3.2.4 Component 334: Non-overlapping Corruption Benchmarker Component

The Non-overlapping Corruption Benchmarker Component functions as a tool for assessing the robustness of neural network models using a benchmark of synthetic corruptions. This tool simulates corruptions similar to natural corruptions and tests their impact on image datasets. It evaluates the accuracy drop caused by these corruptions and measures the impact on the model's performance as the severity of the corruption is modified.

3.2.5 Component 335: Time-series Robustness Characterizer Component

The Time-series Robustness Characterizer Component evaluates the robustness of models on time series data using amplification methods with noise functions. These functions are implemented as Python scripts. Specifically, for the time series use case, a frequency-keyed noise function is provided. Each function includes a sample of noise data and a plot describing the evaluation results.

3.2.6 Component 3141: Chiru

The Chiru component is a tool developed to assess the performance of AI models against input perturbations such as Gaussian noise and Gaussian blur. It provides a graphical interface to visualize the results. Chiru does not inherently support specific model types (e.g., TensorFlow, PyTorch, ONNX, NNET); it is the user's responsibility to add them.

3.2.7 Component 4192: ML watermarking

The ML watermarking component enables black-box watermarking of ML models to reinforce ownership protection. The objective is to investigate how ML watermarking can protect the ownership rights of model creators and ensure traceability of ML models. This component is not focused on evaluating traditional ML models but rather on assessing watermarked models. The component includes three main categories of attacks against watermarking. The first is watermark removal, where an attacker attempts to remove the watermark from

the model. The second is the ambiguity attack, where an attacker casts doubt on legitimate ownership by providing counterfeit watermarks. The third is the evasion attack, where an attacker tries to evade watermark verification, thereby disabling model identification.

An overview of the compatibility of the presented tools with respect to data type and model type is given in tabular Figures 3 and 4, respectively. Note that the column headers in Figures 3 and 4 correspond, respectively, to the number of components as discussed in 3.2.

Data type \ Component	331	332	333	334	335	3141	4192
Images	✘	✘	✘	✘	✘	✘	✘
Tabular	✘	✘	✘	✘	✘	✘	✘
Time-Series	✘	✘	✘	✘	✘	✘	✘
NLP	✘	✘	✘	✘	✘	✘	✘

Figure 3 Supported data types for each component.

Model type \ Component	331	332	333	334	335	3141	4192
Tensorflow	✘	✘	-	✘	✘	-	✘
PyTorch	✘	✘	-	✘	✘	-	✘
ONNX	✘	✘	-	✘	✘	-	✘
NNET	✘	✘	-	✘	✘	-	✘

Figure 4 Supported model types for each component.

4 Conclusions and Future Work

It has been widely argued that ML model robustness is pivotal to reliable AI systems. Accordingly, assessing this robustness is crucial for the successful deployment of ML models in such systems. With this in mind, in the context of the Confiance.ai research program, we proposed a methodological framework to address this need. The framework encompasses engineering processes and a set of supporting tools, providing both methodological and technical support for empirical and formal robustness assessments. However, for simplicity and clarity, this paper focuses solely on the empirical robustness assessment framework.

The proposed framework is part of an end-to-end method for guiding the development of trustworthy ML systems. This method is accessible via a web application known as the "Body of Knowledge", which is currently in development.

The empirical robustness assessment framework presented in this paper (i.e., the process for evaluating ML model robustness by sampling and perturbation and its supporting tools) has been evaluated on industrial use cases such as welding quality inspection, demand forecasting, and visual industrial control. Nevertheless, due to confidentiality reasons on the industrial use cases, the details of the evaluation results cannot be shared publicly.

Yet, an important note to make is that the application of the proposed methodological processes and their associated tools has been highly valuable, as it enabled us to gather feedback from industrial actors and identify areas for improvement. This feedback has already led to actions aimed at enhancing some of the tested tools, such as the maturation of AIMOS based on industrial feedback.

In line with this, as future directions for this research, we plan to pursue two main objectives. First, we will focus on improving the tools that support the engineering processes, using feedback collected from industrial partners. Second, we aim to integrate the developed robustness assessment framework (engineering processes and associated tools) into the "Body of Knowledge", so that all results of our research are consolidated into a single comprehensive reference source.

References

1 ISO/IEC 22989:2022. Information technology — artificial intelligence — artificial intelligence concepts and terminology, 2022.

2 ISO/IEC TR 24029-1. Artificial intelligence (ai)—assessment of the robustness of neural networks—part 1: Overview, 2021.

3 ISO/IEC TR 25059. Iso/iec 25059:2023 – systems and software engineering – systems and software quality requirements and evaluation (square) – quality model for ai-based systems, 2023.

4 Abraham Chan, Arpan Gujarati, Karthik Pattabiraman, and Sathish Gopalakrishnan. The fault in our data stars: studying mitigation techniques against faulty training data in machine learning applications. In *2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, pages 163–171. IEEE, 2022. doi:10.1109/DSN53405.2022.00027.

5 Gregory Chance, Dhaminda B Abeywickrama, Beckett LeClair, Owen Kerr, and Kerstin Eder. Assessing trustworthiness of autonomous systems. *arXiv preprint arXiv:2305.03411*, 2023. doi:10.48550/arXiv.2305.03411.

6 Mahyar Fazlyab, Manfred Morari, and George J Pappas. Probabilistic verification and reachability analysis of neural networks via semidefinite programming. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 2726–2731. IEEE, 2019. doi:10.1109/CDC40024.2019.9029310.

7 Y Guo. Globally robust stability analysis for stochastic cohen–grossberg neural networks with impulse control and time-varying delays. *Ukrainian Mathematical Journal*, 69(8):1049–106, 2017.

8 Chengqiang Huang, Zheng Hu, Xiaowei Huang, and Ke Pei. Statistical certification of acceptable robustness for neural networks. In *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part I 30*, pages 79–90. Springer, 2021. doi:10.1007/978-3-030-86362-3_7.

9 He Huang, Yuzhong Qu, and Han-Xiong Li. Robust stability analysis of switched hopfield neural networks with time-varying delay under uncertainty. *Physics Letters A*, 345(4-6):345–354, 2005.

10 Uyeong Jang, Xi Wu, and Somesh Jha. Objective metrics and gradient descent algorithms for adversarial examples in machine learning. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, pages 262–277, 2017. doi:10.1145/3134600.3134635.

11 Mohd Javaid, Abid Haleem, Ibrahim Haleem Khan, and Rajiv Suman. Understanding the potential applications of artificial intelligence in agriculture sector. *Advanced Agrochem*, 2(1):15–30, 2023.

12 Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *Computer Aided Verification: 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part I 30*, pages 97–117. Springer, 2017. doi:10.1007/978-3-319-63387-9_5.

13 Natan Levy and Guy Katz. Roma: A method for neural network robustness measurement and assessment. In *International Conference on Neural Information Processing*, pages 92–105. Springer, 2022. doi:10.1007/978-981-99-1639-9_8.

14 Ping Li, Fang Xiong, Xibei Huang, and Xiaojun Wen. Construction and optimization of vending machine decision support system based on improved c4. 5 decision tree. *Heliyon*, 10(3), 2024.

15 Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. arXiv:1706.06083.

16 Mohamed-Iliasse Mahraz, Loubna Benabbou, and Abdelaziz Berrado. Machine learning in supply chain management: A systematic literature review. *International Journal of Supply and Operations Management*, 9(4):398–416, 2022.

17 Dietmar PF Möller. Machine learning and deep learning. In *Guide to Cybersecurity in Digital Transformation: Trends, Methods, Technologies, Applications and Best Practices*, pages 347–384. Springer, 2023.

18 Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

19 Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, et al. Adversarial robustness toolbox v1. 0.0. *arXiv preprint arXiv:1807.01069*, 2018.

20 Yixin Nie, Yicheng Wang, and Mohit Bansal. Analyzing compositionality-sensitivity of nli models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6867–6874, 2019. doi:10.1609/AAAI.V33I01.33016867.

21 Wenjie Ruan, Min Wu, Youcheng Sun, Xiaowei Huang, Daniel Kroening, and Marta Kwiatkowska. Global robustness evaluation of deep neural networks with provable guarantees for the hamming distance. *IJCAI-19*, 2019.

22 Jie Wang, Jun Ai, Minyan Lu, Haoran Su, Dan Yu, Yutao Zhang, Junda Zhu, and Jingyu Liu. A survey of neural network robustness assessment in image recognition. *arXiv preprint arXiv:2404.08285*, 2024. doi:10.48550/arXiv.2404.08285.

23 Stefan Webb, Tom Rainforth, Yee Whye Teh, and M Pawan Kumar. A statistical approach to assessing neural network robustness. *arXiv preprint arXiv:1811.07209*, 2018. arXiv:1811.07209.

24 Maurice Weber. *Probabilistic Robustness Guarantees for Machine Learning Systems*. PhD thesis, ETH Zurich, 2023.

25 Lily Weng, Pin-Yu Chen, Lam Nguyen, Mark Squillante, Akhilan Boopathy, Ivan Oseledets, and Luca Daniel. Proven: Verifying robustness of neural networks with a probabilistic approach. In *International Conference on Machine Learning*, pages 6727–6736. PMLR, 2019. URL: http://proceedings.mlr.press/v97/weng19a.html.

26 Matthew Wicker, Luca Laurenti, Andrea Patane, and Marta Kwiatkowska. Probabilistic safety for bayesian neural networks. In *Conference on uncertainty in artificial intelligence*, pages 1198–1207. PMLR, 2020. URL: http://proceedings.mlr.press/v124/wicker20a.html.

27 Cong Xu, Wensheng Chen, Mingkuan Lin, Jianli Lu, Yunghsiao Chung, Jiahui Zou, Ciliang Yang, et al. Applications and challenges of hybrid artificial intelligence in chip age testing: a comprehensive review. *Journal of Artificial Intelligence Practice*, 6(3):70–75, 2023.