
**Artificial Intelligence (AI) —
Assessment of the robustness of
neural networks —**

**Part 1:
Overview**

Copyrighted document, no reproduction or circulation
For review by FG on AI in healthcare
Oct 2024



Copyrighted document, no reproduction or circulation
For review by FG on AI in healthcare
Oct 2024



COPYRIGHT PROTECTED DOCUMENT

© ISO/IEC 2021

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier; Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword	iv
Introduction	v
1 Scope	1
2 Normative references	1
3 Terms and definitions	1
4 Overview of the existing methods to assess the robustness of neural networks	3
4.1 General	3
4.1.1 Robustness concept	3
4.1.2 Typical workflow to assess robustness	3
4.2 Classification of methods	6
5 Statistical methods	7
5.1 General	7
5.2 Robustness metrics available using statistical methods	8
5.2.1 General	8
5.2.2 Examples of performance measures for interpolation	8
5.2.3 Examples of performance measures for classification	9
5.2.4 Other measures	13
5.3 Statistical methods to measure robustness of a neural network	14
5.3.1 General	14
5.3.2 Contrastive measures	14
6 Formal methods	14
6.1 General	14
6.2 Robustness goal achievable using formal methods	15
6.2.1 General	15
6.2.2 Interpolation stability	15
6.2.3 Maximum stable space for perturbation resistance	15
6.3 Conduct the testing using formal methods	16
6.3.1 Using uncertainty analysis to prove interpolation stability	16
6.3.2 Using solver to prove a maximum stable space property	16
6.3.3 Using optimization techniques to prove a maximum stable space property	16
6.3.4 Using abstract interpretation to prove a maximum stable space property	17
7 Empirical methods	17
7.1 General	17
7.2 Field trials	17
7.3 A posteriori testing	18
7.4 Benchmarking of neural networks	19
Annex A (informative) Data perturbation	20
Annex B (informative) Principle of abstract interpretation	25
Bibliography	26

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives or www.iec.ch/members_experts/refdocs).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents) or the IEC list of patent declarations received (see patents.iec.ch).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html. In the IEC, see www.iec.ch/understanding-standards.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 42, *Artificial intelligence*.

A list of all parts in the ISO/IEC 24029 series can be found on the ISO and IEC websites.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html and www.iec.ch/national-committees.

Introduction

When designing an AI system, several properties are often considered desirable, such as robustness, resiliency, reliability, accuracy, safety, security, privacy. A definition of robustness is provided in 3.6. Robustness is a crucial property that poses new challenges in the context of AI systems. For example, in AI systems there are some risks specifically tied to the robustness of AI systems. Understanding these risks is essential for the adoption of AI in many contexts. This document aims at providing an overview of the approaches available to assess these risks, with a particular focus on neural networks, which are heavily used in industry, government and academia.

In many organizations, software validation is an essential part of putting software into production. The objective is to ensure various properties including safety and performance of the software used in all parts of the system. In some domains, the software validation and verification process is also an important part of system certification. For example, in the automotive or aeronautic fields, existing standards, such as ISO 26262 or Reference [2], require some specific actions to justify the design, the implementation and the testing of any piece of embedded software.

The techniques used in AI systems are also subject to validation. However, common techniques used in AI systems pose new challenges that require specific approaches in order to ensure adequate testing and validation.

AI technologies are designed to fulfil various tasks, including interpolation/regression, classification and other tasks.

While many methods exist for validating non-AI systems, they are not always directly applicable to AI systems, and neural networks in particular. Neural network systems represent a specific challenge as they are both hard to explain and sometimes have unexpected behaviour due to their non-linear nature. As a result, alternative approaches are needed.

Methods are categorized into three groups: statistical methods, formal methods and empirical methods. This document provides background on these methods to assess the robustness of neural networks.

It is noted that characterizing the robustness of neural networks is an open area of research, and there are limitations to both testing and validation approaches.

Copyrighted document, no reproduction or circulation
For review by FG on AI in healthcare
Oct 2024

Artificial Intelligence (AI) — Assessment of the robustness of neural networks —

Part 1: Overview

1 Scope

This document provides background about existing methods to assess the robustness of neural networks.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

- ISO Online browsing platform: available at <https://www.iso.org/obp>
- IEC Electropedia: available at <http://www.electropedia.org/>

3.1 artificial intelligence AI

<system>capability of an engineered system to acquire, process and apply knowledge and skills

3.2 field trial

trial of a new system in actual situations for which it is intended (potentially with a restricted user group)

Note 1 to entry: Situation encompasses environment and process of usage.

3.3 input data

data for which a deployed machine learning model calculates a predicted output or inference

Note 1 to entry: Input data is also referred to by machine learning practitioners as out-of-sample data, new data and production data.

3.4

neural network

neural net

NN

artificial neural network

ANN

network of primitive processing elements connected by weighted links with adjustable weights, in which each element produces a value by applying a non-linear function to its input values, and transmits it to other elements or presents it as an output value

Note 1 to entry: Whereas some neural networks are intended to simulate the functioning of neurons in the nervous system, most neural networks are used in artificial intelligence as realizations of the connectionist model.

Note 2 to entry: Examples of non-linear functions are a threshold function, a sigmoid function and a polynomial function.

[SOURCE: ISO/IEC 2382:2015, 2120625, modified — Abbreviated terms have been added under the terms and Notes 3 to 5 to entry have been removed.]

3.5

requirement

statement which translates or expresses a need and its associated constraints and conditions

[SOURCE: ISO/IEC/IEEE 15288:2015, 4.1.37]

3.6

robustness

ability of an AI system to maintain its level of performance under any circumstances

Note 1 to entry: This document mainly describes data input circumstances such as domain change but the definition is broader not to exclude hardware failure and other types of circumstances.

3.7

testing

activity in which a system or component is executed under specified conditions, the results are observed or recorded, and an evaluation is made of some aspect of the system or component

[SOURCE: ISO/IEC/IEEE 26513:2017, 3.42]

3.8

test data

subset of *input data* (3.3) samples used to assess the generalization error of a final machine learning (ML) model selected from a set of candidate ML models

[SOURCE: Reference [2]]

3.9

training dataset

set of samples used to fit a machine learning model

3.10

validation

confirmation, through the provision of objective evidence, that the *requirements* (3.5) for a specific intended use or application have been fulfilled

[SOURCE: ISO/IEC 25000:2014, 4.41, modified — Note 1 to entry has been removed.]

3.11**validation data**

subset of *input data* (3.3) samples used to assess the prediction error of a candidate machine learning model

Note 1 to entry: Machine learning (ML) model *validation* (3.10) can be used for ML model selection.

[SOURCE: Reference [2]]

3.12**verification**

confirmation, through the provision of objective evidence, that specified requirements have been fulfilled

[SOURCE: ISO/IEC 25000:2014, 4.43, modified — Note 1 to entry has been removed.]

4 Overview of the existing methods to assess the robustness of neural networks

4.1 General

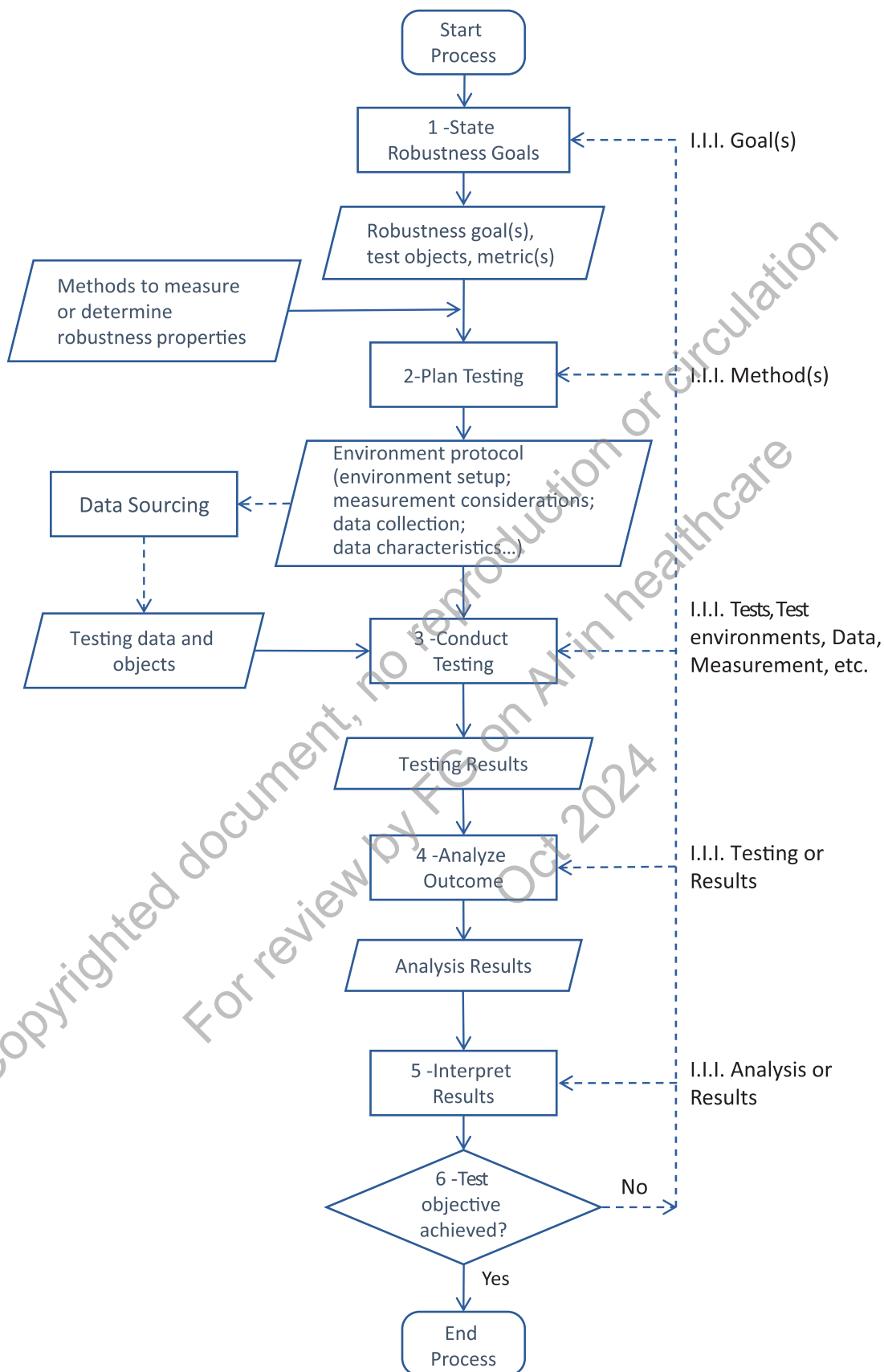
4.1.1 Robustness concept

Robustness goals aim at answering the question “To what degree is the system required to be robust?” or “What are the robustness properties of interest?”. Robustness properties demonstrate the degree to which the system performs with atypical data as opposed to the data expected in typical operations.

4.1.2 Typical workflow to assess robustness

This subclause explains how the robustness of neural networks is assessed for different classes of AI applications such as classification, interpolation and other complex tasks.

There are different ways to assess the robustness of neural networks using objective information. A typical workflow for determining neural network (or other technique) robustness is as shown in [Figure 1](#).



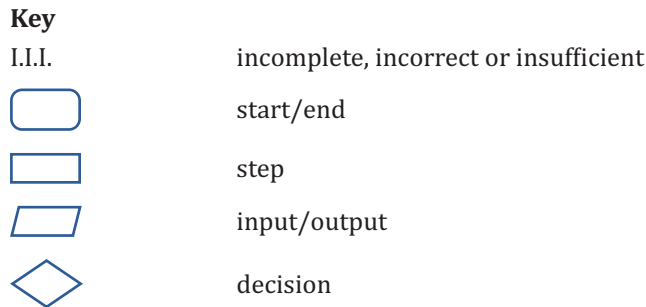


Figure 1 — Typical workflow to determine neural network robustness

Step 1: State robustness goals

The process begins with a statement of the robustness goals. During this initial step, the targets to be tested for robustness are identified. The metrics to quantify the objects that demonstrate the achievement of robustness are subsequently identified. This constitutes the set of decision criteria on robustness properties that can be subject to further approval by relevant stakeholders (see ISO/IEC/IEEE 16085:2021, 7.4.2).

Step 2: Plan testing

This step plans the tests that demonstrate robustness. The tests rely on different methods, for example: statistical, formal or empirical methods. In practice, a combination of methods is used. Statistical approaches usually rely on a mathematical testing process and are able to illustrate a certain level of confidence in the results. Formal methods rely on formal proofs to demonstrate a mathematical property over a domain. Empirical methods rely on experimentation, observation and expert judgement. In planning the testing, the environment setup needs to be identified, data collection planned, and data characteristics defined (that is, which data element ranges and data types will be used, which edge cases will be specified to test robustness, etc.). The output of Step 2 is a testing protocol that comprises a document stating the rationale, objectives, design and proposed analysis, methodology, monitoring, conduct and record-keeping of the tests (more details of the content of a testing protocol are available through the definition of the clinical investigation plan found in ISO 14155:2020, 3.9).

Step 3: Conduct testing

The testing is then conducted according to the defined testing protocol, and outcomes are collected. It is possible to perform the tests using a real-world experiment or a simulation, and potentially a combination of these approaches.

Step 4: Analyze outcome

After completion, tests outcomes are analysed using the metrics chosen in Step 1.

Step 5: Interpret results

The analysis results are then interpreted to inform the decision.

Step 6: Test objective achieved?

A decision on system robustness is then formulated given the criteria identified earlier and the resulting interpretation of the analysis results.

If the test objectives are not met, an analysis of the process is conducted and the process returns to the appropriate preceding step, in order to alleviate deficiencies, e.g. add robustness goals, modify or add metrics, add consideration of different aspects to measure, re-plan tests, etc.

AI systems that significantly rely on neural networks, particularly deep neural networks (DNN), bear built-in malfunctions. These malfunctions are showing up by a system behaviour that resembles

an occurrence of a conventional software. Typical situations have been demonstrated by feeding "adversarial examples" to object recognition systems, e.g. in Reference [5]. These built-in errors of DNNs are not simple to "fix". Research on this problem shows that there are measures to improve the robustness of DNNs with respect to adversarial examples, but this works to a certain degree only [6],[7]. However, if detected during a test procedure, the AI system is able to signal a problem when an associated input pattern is encountered.

Data sourcing:

Data sourcing is the process of selecting, producing and/or generating the testing data and objects that are needed for conducting the testing.

This sometimes includes consideration of legal or other regulatory requirements, as well as practical or technical issues.

The testing protocol contains the requirements and the criteria necessary for data sourcing. Data sourcing issues and methods are not covered in detail in this document.

Especially the following issues can have an impact on robustness:

- scale;
- diversity, representativeness, and range of outliers;
- choice of real or synthetic data;
- datasets used specifically for robustness testing;
- adversarial and other examples that explore hypothetical domain extremes;
- composition of training, testing, and validation datasets.

4.2 Classification of methods

Following the workflow defined above for determining robustness, the remainder of this document describes the methods and metrics applicable to the various testing types, i.e. statistical, formal and empirical methods.

Statistical approaches usually rely on a mathematical testing process on some datasets, and help ensure a certain level of confidence in the results. Formal methods rely on a sound formal proof in order to demonstrate a mathematical property over a domain. Formal methods in this document are not constrained to the traditional notion of syntactic proof methods and include correctness checking methods, such as model checking. Empirical methods rely on experimentation, observation and expert judgement.

While it is possible to characterize a system through either observation or proof, this document chooses to separate observation techniques into statistical and empirical methods. Statistical methods generate reproducible measures of robustness based on specified datasets. Empirical methods produce data that can be analysed with statistical methods but is not necessarily reproducible due to the inclusion of subjective assessment. Therefore, it is usually necessary that methods from both categories be performed jointly.

Thus, this document first considers statistical approaches which are the most common approaches used to assess robustness. They are characterized by a testing approach defined by a methodology using mathematical metrics. This document then examines approaches to attain a formal proof that are increasingly being used to assess robustness. Finally, this document presents empirical approaches that rely on subjective observations that complement the assessment of robustness when statistical and formal approaches are not sufficient or viable.