Objectives and strategies for automated testing of AI-based perception systems in railroad engineering.

Version 1.0
December 2021

Supported by:

Federal Ministry
for Economic Affairs
and Energy

on the basis of a decision
by the German Bundestag

Authors: *Jürgen Grossmann, Johannes Viehmann, Dorian Knoblauch, Lando Petersen, Sami Kharma* (Fraunhofer FOKUS), *Hans-Werner Wiesbrock, Nicolas Grube* (IT-Power Solutions), *Mariele Motta* (Neurocat)

Content

Overview

1. Introduction

The following document describes objectives and strategies for automated testing of AI-based perception systems. The description is made along a set of higher-level objectives that structure the document into individual chapters. The high-level objectives are shown in Figure 1.

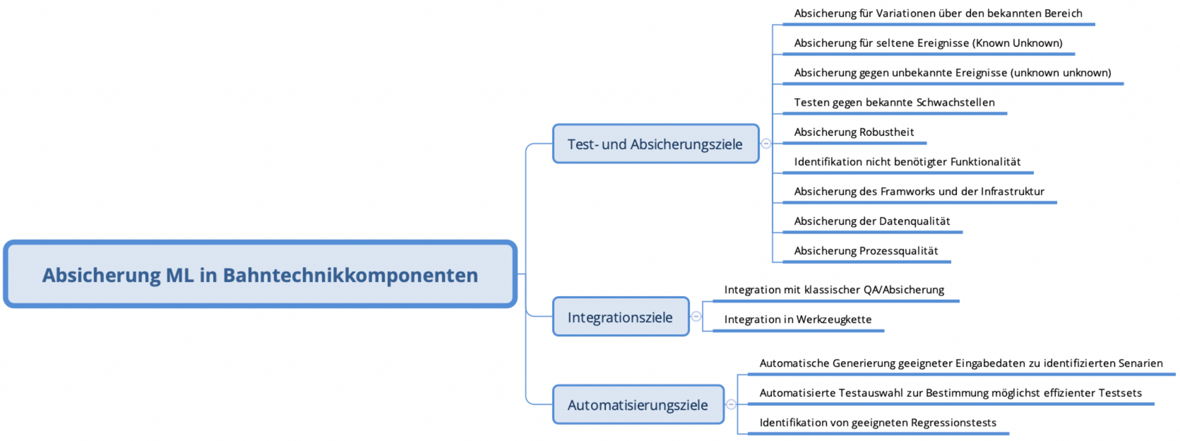
Absicherung ML in Bahntechnikkomponenten
Test- und Absicherungsziele
Absicherung für Variationen über den bekannten Bereich
Absicherung für seltene Ereignisse (Known Unknown)
Absicherung gegen unbekannte Ereignisse (unknown unknown)
Testen gegen bekannte Schwachstellen
Absicherung Robustheit
Identifikation nicht benötigter Funktionalität
Absicherung des Framworks und der Infrastruktur
Absicherung der Datenqualität
Absicherung Prozessqualität
Integrationsziele
Integration mit klassischer QA/Absicherung
Integration in Werkzeugkette
Automatisierungsziele
Automatische Generierung geeigneter Eingabedaten zu identifizierten Senarien
Automatisierte Testauswahl zur Bestimmung möglichst effizienter Testsets
Identifikation von geeigneten Regressionstests

Figure 1 Higher-level objectives

Each high-level objective has a chapter dedicated to it, describing and decomposing it in more detail. The description and decomposition are done along tables. In a first table, each higher-level objective is assigned one or more strategies for achieving the objective. Against the background of the strategy, the higher-level objective is decomposed into further subordinate or derived objectives.

Table 1 Decomposition of the higher-level goals as an example

Objective	**Safeguarding for variations of the known application scope**
Strategy	• Review of training data and checking whether all objects, relevant object combinations and environmental conditions are sufficiently represented in the data. • Dynamic testing on all relevant objects, object combinations and environmental conditions.
Derived objectives	• Identification of relevant objects, object combinations and environmental conditions. • Assignment of relevance and risk criteria to objects, object combinations and environmental conditions. • Identification and development of test metrics and procedures to test data sets for completeness against the real expected environment. • Derivation of abstract test cases for dynamic testing over all relevant objects, object combinations and environmental conditions.

For each derived objective, another table is created that defines strategies, methods and tools, expected results, requirements, and constraints for achieving the derived objective. Boundary conditions can be labelled as pre-conditions (Pre:) and post-conditions (Post:).

Table 2 Presentation of the derived objectives as an example

Derived objective	**Identification of relevant objects, object combinations and environmental conditions.**
Strategy	• Systematic identification o relevant objects (persons, animals, obstacles) o relevant scenes (occluded object, multiple objects nearby, partial view and side/angle view, similar objects at different distances that therefore appear to be different sizes when in fact they belong to the same class, etc.) o relevant weather conditions (e.g., environmental changes such as snow-covered rail, faded signal markers, weather conditions such as snow, fog, and rain) o relevant lighting variations as a function of time of day • Combinations related to invariance requirements such as translational invariance, rotational invariance, colour, and size invariance. • Systematic combination of identified objects, weather conditions, illumination variations.
Method and tools	• Method and tool for systematic modelling of objects and their relationships (ontologies) • Combinatorics to derive scenarios with the help of the ontology
Expected results	• Ontology with the description of relevant objects and their relationships. • Coverage measures to define completeness.
Requirements to fulfil the objectives	• Development of suitable ontologies and modelling methods for discrete modelling of ODD or sub-problems thereof.
Constraints	• Pre: Definition of an ODD with an initial description of all relevant objects and events

2. Test objectives

2.1. Safeguarding/testing the performance for variations of the known application scope

Objective	**Safeguarding/testing the performance for variations of the known application scope**
Strategy	• Review of training data and checking whether all objects, relevant object combinations and environmental conditions are sufficiently represented in the data. • Dynamic testing on all relevant objects, object combinations and environmental conditions.
Derived objectives	• Identification of relevant objects, object combinations and environmental conditions. • Assignment of relevance and risk criteria to objects, object combinations and environmental conditions. • Identification and development of test metrics and procedures to test data sets for completeness against the real expected environment. • Derivation of abstract test cases for dynamic testing over all relevant objects, object combinations and environmental conditions.
Project partner	Fraunhofer FOKUS

Table 3 Fraunhofer FOKUS: *Identification of relevant objects, object combinations and environmental conditions.*

Derived objective	**Identification of relevant objects, object combinations and environmental conditions.**
Strategy	• Systematic identification ○ relevant objects (persons, animals, obstacles) ○ relevant scenes (occluded object, multiple objects nearby, partial view and side/angle view, similar objects at different distances that therefore appear to be different sizes when in fact they belong to the same class, etc.) ○ relevant weather conditions (e.g., environmental changes such as snow-covered rail, faded signal markers, weather conditions such as snow, fog, and rain) ○ relevant lighting variations as a function of time of day • Combinations related to invariance requirements such as translational invariance, rotational invariance, colour and size invariance. • Systematic combination of identified objects, weather conditions, illumination variations.
Method and tools	• Method and tool for systematic modelling of objects and their relationships (ontologies) • Combinatorics to derive scenarios with the help of the ontology
Expected results	• Ontology with the description of relevant objects and their relationships. • Coverage measures to define completeness.
Requirements to fulfil the objectives	• Development of suitable ontologies and modelling methods for discrete modelling of ODD or sub-problems thereof.
Constraints	• Pre: Definition of an ODD with an initial description of all relevant objects and events

Table 4 Fraunhofer FOKUS: *Assignment of relevance and risk criteria to objects, object combinations and environmental conditions.*

Derived objective	**Assignment of relevance and risk criteria to objects, object combinations and environmental conditions.**
Strategy	• Specification of risk factors for individual objects and factors • Specification of risk factors, which are the combination of
Method and tools	• Risk analysis, fault tree analysis, HAZOP
Expected results	• Associated risks to individual objects and factors. • Associated risks to more complex combinations (objects, environmental conditions).
Requirements to fulfil the objectives	• Development of a risk analysis procedure to assign risk factors to individual objects with regard to possible classification errors. • Development of a risk analysis procedure for evaluating more complex combinations of objects and environmental conditions with regard to possible classification errors.
Constraints	

Table 5 Fraunhofer FOKUS: *Identification and development of test metrics and procedures to test data sets for completeness against the real expected environment.*

Derived objective	**Identification and development of test metrics and procedures to test data sets for completeness against the real expected environment.**
Strategy	• Modelling and classification of the real environment along the identified scenarios. • Application of statistical methods to measure distribution differences in the data set and in the real environment.
Method and tools	• Development of a method for analysing the distributions of objects in image data against a given environment model. •
Expected results	• Completeness metrics for data related to identified objects and scenarios
Requirements to fulfil the objectives	• Tools for analysing existing datasets for representativeness against a given model (objects and their distributions).
Constraints	

Table 6 Fraunhofer FOKUS: *Derivation of abstract test cases for dynamic testing over all relevant objects, object combinations and environmental conditions.*

Derived objective	**Derivation of abstract test cases for dynamic testing over all relevant objects, object combinations and environmental conditions.**
Strategy	• Combinatorial testing • Risk based testing
Method and tools	• Development and use of tools and procedures for combinatorial and risk-based testing of perceptual systems in the railroad domain.
Expected results	• Test cases in the form of parameter sets for systematic testing of the input area. • Test oracle for the evaluation of the tests • Coverage measures for defining completeness.
Requirements to fulfil the objectives	• Development and evaluation of procedures for combinatorial testing that are suitable for deriving meaningful parameter sets based on the environmental information. • Provision of procedures for deriving test evaluations (oracles) for the individual parameter sets. • Demonstrate the completeness of the approach with respect to the test objective.
Constraints	• Post: Provide procedures for generating representatives (images) for testing.

2.2. Safeguarding/testing the performance in case of rare or unknown events (known unknown/unknown unknown)

Objective	Evaluation of the DNN Model regarding its' Behaviour in Cases of Rare Events
Strategy	• Analysis of the systems behaviour when inferring on domain-specific test data (possible natural inputs, train/ test split) • Evaluation of this behaviour regarding proportional coverage of the whole systems behaviour • Generation of additional test data to cover a higher proportion of the systems behaviour • Final statistical evaluation of the systems behaviour regarding the correctness when inferring on domain-specific test data and on the additionally generated test data
Derived Objectives	• Finding appropriate methods to evaluate a systems behaviour regarding how much of the possible behaviour is covered by a given set of inputs • Development of methods to generate test data which covers a higher amount of the systems behaviour.

__Table 7 Fraunhofer FOKUS:__ Finding Appropriate Methods to Evaluate a Systems Behavior Regarding How Much of the Possible Behavior is Covered by a Given Set of Inputs

Derived objective	Finding Appropriate Methods to Evaluate a Systems Behaviour Regarding How Much of the Possible Behaviour is Covered by a Given Set of Inputs
Strategy	• Research survey about common evaluation methods and derived metrics • Application of these methods on the concrete use case computer vision systems in locomotive decision systems • Analysis of the derived evaluation metrics regarding relevance in the concrete use cases of the system (preceded risk analysis, certain metrics might over-evaluate risks that are not common in our use case) • Analysis of the evaluation methods regarding reproducibility and explainability
Method and tools	• Research paper from the automotive domain • Open source implementations of common evaluation methods and derived metrics
Expected results	• Comprehensive study of methods and derived metrics in the concrete use case on the properties of relevance, reproducibility and explainability
Requirements to fulfil the objectives	• Comprehensive data set (training set and test set) • Trained computer vision model in the locomotive domain • Risk analysis of possible events in the concrete use case to evaluate the derived metrics regarding their relevance (preceded risk analysis)
Constraints	

__Table 8 Fraunhofer FOKUS__: Development of Methods to Generate Test Data Which Covers a Higher Amount of the Systems Behavior

Derived objective	Development of Methods to Generate Test Data Which Covers a Higher Amount of the Systems Behaviour
Strategy	• Research survey about common test generation methods • Selection of test generation methods based on relevance in the concrete use cases of the system • Application/ implementation of the test generation methods in the application
Method and tools	• Research paper from the automotive domain • Open source implementations of common test generation methods

Expected results	• Generation of test data that lead to higher proportional coverage of the systems behaviour
Requirements to fulfil the objectives	• Metrics to evaluate a systems behaviour regarding the amount of coverage exercised with a given set of inputs • Comprehensive data set (training set and test set) • Trained computer vision model in the locomotive domain
Constraints	

2.3. Safeguarding robustness

__Table 9 ITPower Solutions:__ safeguarding robustness

Objective	**Safeguarding robustness**
Strategy	• Dynamic testing of robustness • Review of training process • Review of training data
Derived Objective	• Identification of relevant perturbations and adversarial attacks (1,2,3)

__Table 10 ITPower Solutions:__ Identification of relevant perturbations and adversarial attacks

Derived Objective	Identification of relevant perturbations and adversarial attacks
Strategy	1. Identification of · Events described in section 3.3 · Perturbations with reference to the environment (e.g. fog, rain) · Perturbations related to the technical environment (e.g. sensor noise, over/under exposure) · Adversarial attacks 2. Determination of quantitative measures of perturbations. 3. Determination of quality criteria for the robustness of the test object. 4. Analytical determination of robustness (e.g., were robust features learned?). 5. Determination of the degree of robustness of the test object.
Methods and tools	Risk-Based-Analysis, XAI, generators of perturbations and adversarial attacks.
Expected results	• Data sets with robustness-relevant perturbations. • Test cases for robustness testing. • Measure (or metrics) for the robustness level of the test object. • Method for comparing the robustness level of different implementations of the test object.
Requirements to fulfil the objectives	• Development of methods for the identification of relevant perturbations from environmental influences. • Development of methods for the identification of intended perturbations (adversarial attacks). • Development and evaluation of methods and procedures for robustness analysis (e.g. XAI for analytical methods). • Development of a method for quantitative comparability of the degree of robustness between different test objects.
Constraints	• Pre: criteria for accepting the robustness property. • Pre: Definition of the labels to be recognized. • Pre: Availability of necessary interfaces between test object and test system.

2.4. Testing against known vulnerabilities

Table 11 Neurocat: *Testing against known vulnerabilities*

Goal	**Testing against known vulnerabilities**
Strategy	• Assess adversarial robustness and generalization with respect to distribution shifts
Derived Objectives	• Assess vulnerability to adversarial attacks through stress tests • Assess change in performance under corruptions (distribution shift), e.g. rain. • Assess reliance on spurious correlations through tests on training data • Verify Robustness through formal verification methods

Table 12 Neurocat: *Assess vulnerability to adversarial attacks through stress tests*

Derived Goal	**Assess vulnerability to adversarial attacks through stress tests**
Strategy	Test robustness to adversarial attacks through ensemble attacks
Methods and tools	• Implementation of various adversarial attacks. • Adversarial attacks are defined by optimization strategy and attacker budget (e.g. perturbation size given a norm, number of iterations). Ex of attack: FGSM with |delta|L2<0.5 with 10 iterations
Expected results	Systematic evaluation of performance for relevant metrics, attacked classes and attack. E.g. table with index=attack, columns= classes, values= classification accuracy
Requirements for the implementation of the goal	
Requirements for the test process/ environment	• Define set of attacks (attack space is infinite) • Define relevant metrics • Define relevant classes

Table 13 Neurocat: *Assess change in performance under corruptions (distribution shift).*

Derived Goal	**Assess change in performance under corruptions (distribution shift).**
Strategy	Test robustness to distribution shifts
Methods and tools	• Implementation of various corruptions attacks/distribution shifts. E.g. adding noise, or simulating weather conditions. • Corruption attacks with noise can be described by properties of distribution, e.g. Normal distribution with mean=m, variance=v • Distribution shifts such as rain have to be described by some metric/parameter – open problem
Expected results	Systematic evaluation of performance for relevant metrics, attacked classes and attack. E.g. table with index=attack, columns= classes, values= classification accuracy
Requirements for the implementation of the goal	
Requirements for the test process/ environment	• Define set of corruption attacks • Define set of distribution shifts • Define relevant metrics • Define relevant classes

Table 14 Neurocat: *Assess reliance on spurious correlations through tests on training data.*

Derived Goal	**Assess reliance on spurious correlations through tests on training data**
Strategy	Test if model relies on spurious features for task and whether backdoors were introduced during training

Methods and tools	• Detect 'Clever Hans' or 'Backdoors' using Explanation methods or inner layer representations. • Detectors run on samples of the training data. • Human supervision is required to adjust detectors
Expected results	• Estimation of % of training data with Clever Hans. This can be presented as percentage of minority and majority groups. • Detection of Backdoor attack
Requirements for the implementation of the goal	
Requirements for the test process/ environment	• Define set of explanation methods • Define set of detectors • Define thresholds for detectors (specific for used model + training data)

__Table 15 Neurocat:__ Verify Robustness through formal verification methods.

Derived Goal	**Verify Robustness through formal verification methods**
Strategy	• Compute certified bounds for chosen architectures. • Test bounds with adversarial attacks
Methods and tools	• Select verification method(s), e.g. CROWN, FROWN • Select corresponding adversarial attacks • Chose number of data points
Expected results	• Report with performance metrics • Reachability certification (e.g. average interval lenght, average upper reach, lower reach, maximal reach)
Requirements for the implementation of the goal	
Requirements for the test process/ environment	• Relevant architectural choice, understanding/choice of suitable verification method and corresponding adversarial threat

2.5. Safeguarding the quality of the data

__Table 16 ITPower Solutions:__ Safeguarding the quality of the data

Objective	**Safeguarding the quality of the data**
Strategy	Review of learning and test data sets
Derived Objective	• Evaluation of statistical quality characteristics • Evaluation of content-related and technical quality features

__Table 17 ITPower Solutions:__ Evaluation of statistical quality characteristics

Derived Objective	**Evaluation of statistical quality characteristics**
Strategy	1. Assessment of the size and representativeness of the data sets of the learning process (training, validation and test data).. · Description of the data sets using descriptive statistics (e.g., number of objects to be determined such as trees or people and scenarios such as day or night). · Generation of measures of extent and representativeness from characteristics determined above.

	· Evaluation of the extent and representativeness on the basis of given acceptance criteria. 2. Evaluation of completeness of test data sets and consideration of edge cases. · Identification of edge cases by means of risk analysis. · Description of data sets using descriptive statistics. · Evaluation of the completeness and coverage of edge cases on the basis of given acceptance criteria and resulting missing edge cases, if any.
Methods and tools	1. and 2. descriptive statistics, CTE, risk analysis (edge cases).
Expected results	• Procedures to examine data sets for comprehensiveness, completeness, and representativeness. • Procedures for identifying gaps and underrepresented edge cases.
Requirements to fulfil the objectives	• Develop methods for qualitatively assessing the extent, completeness, and representativeness of datasets. • Develop statistical measures to assess data quality.
Constraints	• Pre: delivery of learning process data. • Pre: ODD/use cases (narrative specifications). • Pre: Acceptance criteria for extent, representativeness, completeness, and coverage of edge cases.

***Table 18 ITPower Solutions:** Evaluation of content-related and technical quality features*

Derived Objective	**Evaluation of content-related and technical quality features**
Strategy	1. Evaluation of the correctness and quality of the label of the data. · Selection of samples. · Checking the correctness and quality of the labels of the data. · Derivation of parameters for data quality. 2. Determination of content and technical data quality. · Selection of samples. · Examination of content (e.g. relevance of the depicted scenarios) and technical (negative example e.g. blurred or noisy images) data quality. · Transfer to baseline (data set).
Methods and tools	1. and 2. statistics, 2. methods and tools to ensure the correctness of the labels.
Expected results	• Measure (or metrics) of content-related and technical data quality. • Process for identifying gaps and underrepresented edge cases.
Requirements to fulfil the objectives	• Development of statistical measures to assess data quality. • Development of methods of ensuring the correctness of labels. • Development of a method for quantitative evaluation of technical data quality.
Constraints	• Pre: Delivery of the data of the learning process. • Pre: ODD/ Use cases (narrative specifications).

2.6. Testing for dependence on spurious correlations

***Table 19 ITPower Solutions:** Testing for dependence on spurious correlations*

Objective	Testing for dependence on spurious correlations
Strategy	• Exploratory testing
Derived Objective	• Correlation analysis

Table 20 ITPower Solutions: *Correlation analysis*

Derived Objective	**Correlation analysis**
Strategy	1. Determine pairwise correlation coefficients between features and classifications (e.g., using correlation coefficients, truth matrices, or saliency maps). 2. Evaluation of the correlations. 3. Checking whether acceptance criteria are met.
Methods and tools	Statistical methods, XAI
Expected results	• Method for checking misclassification between features of input data and classification of the output.
Requirements to fulfil the objectives	• Development of methods for testing correlations between input and output data of an ML model.
Constraints	• Pre: Delivery of the ML model to be tested and the input and output data. • Pre: Acceptance criteria for correlation measures.

2.7. Testing for performing at least on human level

Table 21 Fraunhofer FOKUS: *Testing for performing at least on human level*

Objective	**Testing for performing at least on human level**
Strategy	• Compare AI with human train drivers by testing both with the same input data
Derived objectives	• Test AI with real-world train camera data as input and real-world train driver reactions in regular operation as expected output • Test human beings with real-world train camera data in an explicit test situation • Test human beings and AI with artificial generated test data
Project partner	Fraunhofer FOKUS

Table 22 Fraunhofer FOKUS: *Test AI with real-world train camera data as input and real-world train driver reactions in regular operation as expected output*

Derived objective	**Test AI with real-world train camera data as input and real-world train driver reactions in regular operation as expected output**
Strategy	• Use observation in regular train operation
Method and tools	• On train: record front view • Log train driver reactions • Identify critical situations worth testing the AI against with appropriate tool and execute these tests against AI
Expected results	• Shows how AI performs compared to human beings in certain rea-world scenarios, in regular operation (i.e. nothing happens for long times)
Requirements to fulfil the objectives	• Cameras on train in regular operation • Logging system for train driver reactions on train in regular operation • It must be ensured that it is legal to use data recorded on train in regular operation for testing • Tool for automated real-world test data selection and test execution needs to be developed
Constraints	Requires a lot of observation in regular train operation to get a high coverage of what could eventually happen

Table 23 Fraunhofer FOKUS: *Test human beings with real-world train camera data in an explicit test situation*

Derived objective	**Test human beings with real-world train camera data in an explicit test situation**
Strategy	• Use only selected critical situations form recorded real-world data and test human beings with that in an explicit test situation

	- Compare results with reactions of train drivers in the regular operation when the video data was recorded
Method and tools	- Display recorded real-world data of selected critical situations
Expected results	- Get the average latency of human beings under boring regular operation conditions in comparison to test situations where a lot happens all the time
Requirements to fulfil the objectives	- Same as described for “Test AI with real-world train camera data as input and real-world train driver reactions in regular operation as expected output”
Constraints	

Table 24 Fraunhofer FOKUS: *Test human beings and AI with artificial generated test data*

Derived objective	**Test human beings and AI with artificial generated test data**
Strategy	- Create test data for rare situations by using simulation - Test human beings in explicit testing situation and add average latency for real-world regular operation - Test AI - Compare results
Method and tools	- 3D Simulation, critical scenario catalogues
Expected results	- Realistic performance comparison for
Requirements to fulfil the objectives	- Latency for human beings in real-world regular operation as described in “Test human beings with real-world train camera data in an explicit test situation”
Constraints	

3. Automation objectives

3.1. Automated generation of appropriate input data for identified scenarios

Table 25 Fraunhofer FOKUS: Automated generation of appropriate input data for identified scenarios

Objective	Automated generation of appropriate input data for identified scenarios
Strategy	• Simulation-based generation of input data • Generative-based generation of input data
Derived Objectives	• identification of base scenarios or inputs with known-ground truth • generative modifications of inputs with know-ground truth (using e.g. metamorphic relations) • simulation-based test scenario generation (e.g. by game engines)

Table 26 Fraunhofer FOKUS: Simulation-based generation of input data

Objective	Simulation-based generation of input data
Strategy	• Use scenario descriptions as input to a 3D generator or existing scenarios • 3D generator renders the decried scenario with multiple variations and reasonable limits of freedom • Replay the generated scene and capture rendered images and corresponding segmentation maps
Method and Tools	• 3D Rendering, Segmentation of 3D Renderings • Extraction of render-time data from the GPU • Analysis of rendering pipeline states of existing simulators to generate segmentation maps for existing scenarios
Expected results	• Set of (not necessarily photorealistic) Images with corresponding ground truth
Requirements to fulfil the objectives	• A comprehensive 3D generator for the desired scenarios or an existing 3D environment. • Descriptions of scenarios
Constraints	• pre: description of scenarios, existing generator or enviorment • post: procedures for generating representatives (images) for testing.

Table 27 Fraunhofer FOKUS: Generative-based generation of input data

Objective	Generative-based generation of input data
Strategy	1. Pre-process input data properly 2. Design and train a GAN-style architecture
Method and Tools	Machine learning, specifically GAN, WS-GAN, potentially (variational) autoencoders
Expected results	• Variant of a GAN network capable of synthesizing low resolution images which appear to belong to the same dataset as pre-existing input data • Method to synthesize such images using the network and ability to retrain the network on other similarly structured input data
Requirements to fulfil the objectives	• Avoiding (partial) mode-collapse of the GAN network during training • Fine tuning hyperparameters and model architecture to ensure the above holds • Proper application of classical and non-classical image augmentation approaches to fully make use of existing input data
Constraints	• pre: presence of a varied and sufficient numbers of input data, presence of computational resources sufficient for training variants of GAN networks • post: procedures for generating representatives (images) for testing, images likely low resolution

3.2. Automated test selection to determine the most efficient test sets possible

Table 28 Fraunhofer FOKUS*: Automated test selection to determine the most efficient test sets possible*

Objective	**Automated test selection to determine the most efficient test sets possible**
Strategy	• risk-based testing
Derived Objectives	• identification of relevant risk factors • assignment of risk factors to individual tests or test sets • risk-based selection of tests

Table 29 Fraunhofer FOKUS*: Risk-based selection of tests*

Objective	**Risk-based selection of tests**
Strategy	1. identification of relevant objects and scenarios as parameters for the test (ontologies). 2. assignment of risk indicators to the objects and scenarios. 3. derivation of meaningful test oracles 4. selection of tests according to the assigned risk metrics
Method and Tools	Risk analysis and risk-based testing
Expected results	• ontology with the description of relevant objects and their relationships. • risk assessment of the objects and their relationships. • test cases in the form of parameter sets for systematic testing of the input domain. • evaluation of tests in terms of their potential to confirm, mitigate or identify new risks. • test oracle for the evaluation of the tests. • coverage measures to define completeness, taking into account the risks associated with the test.
Requirements to fulfil the objectives	• development of appropriate ontologies and modeling methods for discrete modeling of ODD or sub-problems thereof. • development and evaluation of risk analysis techniques and procedures suitable to differentiate scenarios along the associated risk. • development of a procedure for risk-based selection of tests with the goal of: ○ achieve high test coverage for high-risk scenarios. ○ target tests to reduce existing uncertainties (testing scenarios for which the probability of a correct system response is unknown). ○ testing based on failure heuristics • demonstration of the completeness of the approach with respect to the test objective.
Constraints	• pre: definition of an ODD with a comprehensible description of all relevant objects and events. • pre: risk analysis on the objects and events. • post: procedures for generating representatives (images) for testing.

3.3. Automated identification of test oracles (ground truth)

Table 30 ITPower Solutions: *Automated identification of test oracles (ground truth)*

Objective	**Automated identification of test oracles (ground truth)**
Strategy	• Metamorphic test method for ML solutions
Derived Objective	• Generation of a test oracle by means of metamorphic test methods

Table 31 ITPower Solutions: *Generation of a test oracle by means of metamorphic test methods*

Derived Objective	**Generation of a test oracle by means of metamorphic test methods**

Strategy	2. Qualitative description of logical relations between inputs and outputs of the test object. · Analysis of input and output patterns. · Deduction of metamorphic relations. · Automation of the analysis and generation of metamorphic relations (rule based or ML based). 3. Checking the validity of the metamorphic relations in the test case.
Methods and tools	Methods for the generation of metamorphic relations
Expected results	• Method for automated generation of test oracles. • Tool for (partially) automated generation of test oracles.
Requirements to fulfil the objectives	• Development of methods for the generation of metamorphic relations. • Development of a tool for the (partially) automated generation of test oracles.
Constraints	• Pre: Supply of input and output data.

4. References

[1] Zhang, J. M., Harman, M., Ma, L. & Liu, Y. Machine Learning Testing: Survey, Landscapes and Horizons. arXiv:1906.10742 [cs, stat] (2019).

[2] Humbatova, N. et al. Taxonomy of real faults in deep learning systems. in Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering 1110–1121 (ACM, 2020). doi:10.1145/3377811.3380395.

[3] Poddey, A., Brade, T., Stellet, J. E. & Branz, W. On the validation of complex systems operating in open contexts. arXiv:1902.10517 [cs] (2019).