IVVES

Industrial-grade Verification and Validation of Evolving Systems

Labelled in ITEA3, a EUREKA cluster, Call 5

ITEA3 Project Number 18022

D5.3 – Initial IVVES Methodology for V&V of ES

Due date of deliverable: September 30, 2021
Actual date of submission: October 15, 2021

Start date of project: 1 October 2019   **Duration:** 39 months

Organisation name of lead contractor for this deliverable: OU

Author(s):	Juan Leandro, David Caubilla, Pekka Aho, Tanja E.J. Vos
Status:	ToC
Version number:	1.0
Submission Date:	15-10-2021
Doc reference:	IVVES_Deliverable_D5.3_V1.0.docx
Work Pack./ Task:	WP5 / Task 5.2
Description: *(max 5 lines)*	The IVVES Integrated Methodology describes how one can combine results from WP2, WP3, and WP4 to form a common process to safeguard ES.

Nature:	**R**=Report		
Dissemination Level:	**P U**	Public	**X**
	P P	Restricted to other programme participants	
	R E	Restricted to a group specified by the consortium	
	C O	Confidential, only for members of the consortium	

DOCUMENT HISTORY

Release	Date	Reason of change	Status	Distribution
V0.1	03/09/2021	First draft	Draft	WP5
V1.0	15/10/2021	Final version	Final	ITEA

Table of Contents

Glossary

Abbreviation / acronym	Description
IVVES artifact	An outcome of IVVES project. Types of artifacts we will have are: tools, data, courses, tutorials, tool chains, use cases and success stories.

1. Executive Summary

The IVVES framework is an online platform that acts as a front-end for the public IVVES artifacts to the people outside of the consortium. The aim of the framework is that all the public IVVES outcomes come with a well-defined and clear package of training and possibilities to try them out (experiment with them). Besides the framework, the "IVVES Integrated Methodology for V&V of ES" aims to describe how the results from WP2, WP3 and WP4 can be combined to form a common process. A first approach to this methodology is described in this document. We propose to add a view of artifacts in a domain-specific fashion. That is, a view of domain-specific techniques; a reference domain-focused V&V toolchain that could guide third parties to set up a specific learning (and experimentation) path.

2. Introduction

A first approach to this methodology is described in this document. We propose to add a view of artifacts in a domain-specific fashion. That is, a view of domain-specific techniques; a reference domain-focused V&V toolchain that could guide third parties to set up a specific learning (and experimentation) path.

The document is showing the libraries, tools, components, techniques from (WP2, WP3, WP4) that have been tested to solve the V&V challenges and requirements.

For every domain:

- Cyber-security
- Transportation
- Healthcare
- Finance
- Industrial Automation

a brief description of requirements is provided, following a description of the components that are being integrated (or tested) as par of the verification and validation workflow.

3. IVVES for Cyber-security (F-SECURE)

3.1 Description of requirements and challenges.

The requirements and challenges identified in this domain, are related to the use cases described in "D1.1 – Detailed Description of Requirements from Use Cases". The components, techniques, tools and libraries designed and/or improved and integrated at this stage are covering three sections:

- Security events analysis.
- Test selection and prioritization analysis.
- Automated tests generation and application model inference.

3.2 Description of the Methodology for V&V of ES in this domain.

3.2.1 Security events analysis

DENDRIFT: A DRIFT-AWARE ALGORITHM FOR HOST PROFILING

DenDrift is a tool that can detect three prevalent forms of concept drift and improve the quality of clustering algorithms. Can effectively distinguish between outliers and drifts and is supplied with a concept drift simulator:

- Based on Gaussian mixture models
- Can simulate three forms of drift with user defined magnitudes and durations
- Allows users to adjust the number of clusters before and after the simulated drift

DenDrift has been tested as part of the V&V process for "FS2: Quality assessment and reproducibility of unsupervised ML models", specifically for Model Quality (T2.1).

Reference to component or artifact in IVVES framework (if available).

https://gitlab.com/itea-ivves/wp2/concept-drift/-/blob/master/paper.pdf

COMMAND LINE QUALITY

Computer vulnerabilities can be exploited in a variety of ways. Malicious actors may use a specific exploit, a secret pathway to enter a computer system, or a misconfiguration in one of the system components. In most of these attacks, malicious actors run malicious programs through shell commands. One way to detect malicious activities on a machine is by analyzing the structure of shell commands. "Command Line Quality" is a hybrid approach that combines rule-based systems (based on expert opinions) and Machine Learning to solve the problem. That is, given a new shell command, Command Line Quality finds other similar shell commands. This helps in creating clusters of similar commands and classifying them as safe or malicious. "Command Line Quality" has been tested as part of the V&V process for "FS3: Utilization of complex textual data for machine learning applications", specifically for Training Data Quality (T2.2).

Reference to component or artifact in IVVES framework (if available).

NA

DATA QUALITY WRAPPER

Data Quality Wrapper is a tool that includes tabular statistical verification methods and NLP techniques for assessing quality of tabular and unstructured text data. This solution addresses the challenge of automating the pre-processing step in structured and unstructured (data, together with data visualization step for both and pre-processing of text data It is of crucial importance to understand the underlying assumptions of data in order to prepare it for training an AI model.

"Data Quality Wrapper" has been tested as part of the V&V process for "FS3: Utilization of complex textual data for machine learning applications", specifically for Training Data Quality (T2.2). The DATA QUALITY WRAPPER is a general-purpose tool, and is being tested in several domains. The results will be provided in the next version of the document (D5.4).

Reference to component or artifact in IVVES framework (if available).

https://share.streamlit.io/soft-nougat/dqw-ivves/app.py (live demo)
https://gitlab.com/itea-ivves/wp2/data-quality-wrapper (closed repo)

3.2.2 Test selection and prioritization analysis

PYTEST-RTS

It is a Coverage-based Regression Test Selection (RTS) plugin for PYTEST. The plugin is supposed to be used to execute tests related to changes done locally on developer's machine and in CI environment to test pull requests. It has have been tested as part of the V&V process; improving performance and integrated into CI/CD pipeline for "FS4: Optimization of tests execution", being related to ML-driven Testing (task 3.1).

Reference to component or artifact in IVVES framework (if available).

https://learn.ivves.eu/course/view.php?id=6
https://github.com/F-Secure/pytest-rts
https://ieeexplore.ieee.org/document/9425967

FLAKY TESTS DETECTION

It is a Github actions plugin to check flakiness of tests by calculating fliprates. It has been tested as part of the V&V process for "FS4: Optimization of tests execution".

Reference to component or artifact in IVVES framework (if available).

https://github.com/F-Secure/flaky-tests-detection (used in:
https://github.com/MarketSquare/robotframework-browser in builds)

TESTARCHIVER

TESTARCHIVER is a tool for archiving test results to a SQL database. Improvements and additional support have been tested as part of the V&V process for "FS4: Optimization of tests execution", being related to ML-driven Testing (task 3.1).

Reference to component or artifact in IVVES framework (if available).

https://github.com/salabs/TestArchiver

CHANGEENGINE

CHANGEENGINE is a tool for prioritizing test runs based on previous results. It works by learning which test cases are the most sensitive to changes that are made between test executions. It has been tested, together with TESTARCHIVER, as part of the V&V process for "FS4: Optimization of tests execution".

Reference to component or artifact in IVVES framework (if available).

https://github.com/salabs/ChangeEngine

CODE QUALITY PLUGIN

CODE QUALITTY PLUGIN uses machine learning to predict risky code areas by if a commit contains code that could result in defect. It has been tested as part of the V&V process for "FS4: Optimization of tests execution", being related to ML-driven Testing (task 3.1), investigating data sources to detect code quality with machine learning and also related to Testing with Uncertainties (task 3.2), testing risky pieces of code and predicting code quality defects. The plugin is a general-purpose tool, and is being tested in several domains. The results will be provided in the next version of the document (D5.4).

Reference to component or artifact in IVVES framework (if available).

https://gitlab.com/itea ivves/wp3/code quality

3.2.3 Automated tests generation and application model inference

TESTAR

TESTAR is an open-source tool for scriptless test automation through graphical user interface (GUI). It can be used for monitoring too, especially with state model inference and analysis. The idea is to infer state models by observing the GUI of the system under test during automated testing. The models are inferred from consequent version of the same system, and changes between versions can be detected by comparing the models. TESTAR has been tested as part of the V&V process for "FS5: Automated test generation", being related to ML-driven Testing (task 3.1) and Online Testing & Monitoring (task 3.3), especially with state model inference and analysis.

Reference to component or artifact in IVVES framework (if available).

https://learn.ivves.eu/course/view.php?id=2
https://testar.org
https://github.com/TESTARtool/TESTAR_dev/

4. IVVES for Transportation (ALSTOM)

4.1 Description of requirements and challenges.

Traditional predictive maintenance approaches are based on data collection during system operations. Alstom aim at finding and resolving issues earlier by exploring a new approach with more proactive analytics, complementing the collection of operational data with data from hardware/software development.

4.2 Description of the Methodology for V&V of ES in this domain.

4.2.1 Data Collection and Analytics in Engineering

RELOAD (by RISE)

RELOAD is an automated machine-learning assisted load testing tool that can generate effective and efficient workloads to identify performance bottlenecks. It is being tested as part of the V&V process for "UC2-Transportation: Data Collection and Analytics in Engineering and Operations", for the Next Gen Control Computer Unit.

Reference to component or artifact in IVVES framework (if available).

https://github.com/mahshidhelali/RL-Assisted-Performance-Testing

SMARTOPS TOOL

The tool enables test managers and product managers to have visibility over their DevOps pipeline. The tool collects data from various sources, analyses the data and visualizes it in a dashboard. This enables a data-driven engineering environment.

Reference to component or artifact in IVVES framework (if available).

https://gitlab.com/itea-ivves/wp4/smartops

4.2.2 Data Collection and Analysis in Operations

DATA CLEAN-UP TECHNIQUE

This activity has been performed on AddTrack. The AddTrack system is a railway operational data system owned, developed and maintained by Addiva AB. The activity is part of the V&V process for "UC2-Transportation: Data Collection and Analytics in Engineering and Operations".

Reference to component or artifact in IVVES framework (if available).

"Predictive Failure Data Analysis in the AddTrack"

5. IVVES for Healthcare (PHILIPS)

5.1 Description of requirements and challenges.

The main objective is to improve effectiveness and efficiency of the V&V activities. Current challenges to address are:

- Automation of test case creation based on real customer use. Through analysis of log files, we try to extract the operational workflows that represent most common customer use cases including the typical scan protocols used with these workflows.

- For training the algorithms and for verification activities we need clinical data sets. The amount of clinical data needed, limitations with scanning on 'volunteers' or using hospital data and privacy rules restrict availability and possibility in using real clinical data and therefore we're exploring possibilities in use of synthetic data.

- Our platform concept and following an evolutionary/incremental development strategy requires tests to be repeated/executed on different HW configurations and increases the level of regression testing with each development cycle. How can we use AI for smart selection of (regression) tests across applicable HW configurations?

- Our defect management process has several decision-making steps. Can AI support and as such increase efficiency of the decision-making process for defects?

5.2 Description of the Methodology for V&V of ES in this domain.

OPERATIONAL PROFILE DRIVEN TEST FRAMEWORK

The OPERATIONAL PROFILE DRIVEN TEST FRAMEWORK being developed is a proprietary Philips tool, tuned for the MR product, to extract and execute operational workflows. The process the tool supports will be shared with and can be applied in any of the IVVES domains.

3D GAN ALGORITHMS

Currently, 3D GAN algorithms are being experimented to generate 3D MR images of the brain.

Reference to component or artifact in IVVES framework (if available).

https://github.com/neuronets/progressivegan3d
https://arxiv.org/abs/1710.10196

CHANGEENGINE

CHANGEENGINE is a tool for prioritizing test runs based on previous results. It works by learning which test cases are the most sensitive to changes made between test executions. Collaboration with F-Secure will be setup to share and accelerate learnings

Reference to component or artifact in IVVES framework (if available).

https://github.com/salabs/ChangeEngine

6. IVVES for Finance

6.1 Description of requirements and challenges.

6.1.1 Improved V&V techniques with AI (ING)

The main is to improve the V&V techniques through the entire development and testing lifecycle of the ING applications (mobile and web) for individual customers by replacing manual work with AI-based V&V tools in the development and testing lifecycle. For example: facilitate error detection and correction in a new release candidate and advise on go-live.

6.1.2 V&V for ESG investment system (CCTL/NETCHECK)

The requirements and challenges identified in this domain, are related to the use cases described in "D1.1 – Detailed Description of Requirements from Use Cases". The main goal is to provide an effective tool to assess and validate performance and reliability of ML-enabled systems for ESG scoring based on new AI-based V&V techniques; increasing the adoption of ML-based services from Home Offices, Brokerage Agencies and Fintech companies.

6.2 Description of the Methodology for V&V of ES in this domain.

6.2.1 Improved V&V techniques with AI (ING)

What libraries, tools, components, techniques… are being combined (WP2,WP3,WP4) to solve the V&V challenges and requirements. If any of the component is present as artifact, please add the reference.

TESTAR

TESTAR is an open-source tool for scriptless test automation through graphical user interface (GUI). TESTAR has been tested as part of the V&V process for "UC4- Finance domain: Improved V&V techniques with AI", aiming to provide improved ML algorithms for action selection, and using state model inference for automated change detection. TESTAR tool has been extended to support testing of mobile applications and evaluated by using it for testing ING mobile banking app. The early results indicate that TESTAR complements the existing scripted GUI testing by increasing the combined coverage.

Reference to component or artifact in IVVES framework (if available).

https://learn.ivves.eu/course/view.php?id=2
https://testar.org
https://github.com/TESTARtool/TESTAR_dev

6.2.2 V&V for ESG investment system (CCTL/NETCHECK)

What libraries, tools, components, techniques… are being combined (WP2,WP3,WP4) to solve the V&V challenges and requirements. If any of the component is present as artifact, please add the reference.

GROUND-TRUTH MANAGER

This component is in charge of ingesting information related to other sources (i.e. ESG Data from Reuters) and provides a HIL (Human in the Loop) interface to confirm and/or enhance data labelling. It is being tested as part of the V&V process for "UC4 – Finance: ESPFIN1: V&V for ESG investment system".

Reference to component or artifact in IVVES framework (if available).

https://github.com/arielmoretti/ESPFIN-ESG-IVVES

ESP-FIN SERVER

This component is providing a set of features that are being tested as part of the V&V process for "UC4 – Finance: ESPFIN1: V&V for ESG investment system". The server is offering an ESG Topic Extractor, an ESG Scorer and the XAI Provider. This last component is providing, for a given asset, a summary of the main sources (news, reports…) impacting any update on the ESG scoring. It is being tested as part of the V&V process for "UC4 – Finance: ESPFIN1: V&V for ESG investment system".

Reference to component or artifact in IVVES framework (if available).

https://github.com/arielmoretti/ESPFIN-ESG-IVVES

7. IVVES for Industrial Automation

7.1 Description of requirements and challenges.

7.1.1 System Level Regression Test Analysis (ABB)

The ABB robot controller software is a long-lived evolving system with its roots in the early 90s. It's a one-size-fits-all open system which can be customized to fulfil a lot of different customer cases, such as picking, arcwelding, spot welding, painting, etc. The robot controller software has several configurations, and all different configurations make it hard to measure and predict how the hardware/software resources are utilized. When a problem arises, it is also hard to tell in which software build the problem was introduced.

In IVVES, we would like to measure performance and system utilization in each new daily build of the robot controller software. Analyse the deviations between builds. Decide if the deviations are within normal fluctuation or if they are suspicious.

Performance testing at ABB is currently a manual activity which is done after functional testing and before releasing a new release of the application. The activity is labor intensive and it is hard to measure and predict how the hardware/software resources are utilized. When a problem arises, it is also hard to tell in which software build the problem was introduced.

7.2 Description of the Methodology for V&V of ES in this domain.

7.2.1 System Level Regression Test Analysis (ABB)

The solutions that are being developed for the ABB's use-case are based on the algorithms and techniques implemented in the following tools:

RELOAD (by RISE)

RELOAD is an automated machine-learning assisted load testing tool that can generate effective and efficient workloads to identify performance bottlenecks.

Reference to component or artifact in IVVES framework (if available).

https://github.com/mahshidhelali/RL-Assisted-Performance-Testing

SaFReL (by RISE)

SAFREL is a self-adaptive fuzzy reinforcement learning performance test agent generating platform-based test cases. It learns how to tune the resource availability to reach the intended performance breaking point for different types of SUTs (in terms of their sensitivity to resources).

Reference to component or artifact in IVVES framework (if available).

https://github.com/mahshidhelali/RL-Assisted-Performance-Testing
https://learn.ivves.eu/course/view.php?id=4

ACTA

ACTA is an automated test generation method called ACTA for black-box performance testing. ACTA is based on active learning, which means that it does not require a large set of historical test data to learn about the performance characteristics of the system under test. Instead, it dynamically chooses the tests to execute using uncertainty sampling. It has been tested as part of the V&V process for "UC5-Industrial Automation: System Level Regression Test Analysis", being related to ML-driven Testing (task 3.1), implementing a deep learning algorithm for performance test generation.

Reference to component or artifact in IVVES framework (if available).

https://arxiv.org/abs/2104.02102

8. IVVES General Purpose Solutions

8.1 MACAU

Currently, MACAU is experimented with in a number of de-facto standard toy cases, equities ranking in **financial** domain, energy demand prediction as a part of a hybrid model solution, synthesis pathway analysis in synthetic biology, to name a few, in addition to we being open for interesting IVVES challenges So far, the method has worked exactly as expected out of box which can be sometimes abnormal in the realm of machine learning Especially uncertainty modelling seems to be promising.

8.2 XAI Toolkit: Glassbox

This solution addresses the challenge of transparency in black box AI models, specifically understanding the model's sequential decisions by creating hierarchical explanations. The current approach, focused on image data, could be applied to a wide range of solutions.

8.3 Deeper: Simulation based System level Testing of ML Systems

This solution is a simulation based test generator that uses an evolutionary procedure for generating critical (system level test cases to test a ML system in an automotive domain, i e a lane keeping system

8.4 Testing the Learning Algorithm

An oracle centred approach to evaluate (the rightness of) implementations of learning algorithms for decision trees.

8.5 MLOps Pipeline

The two main objectives in this case are:
- Augmenting DevOps (SW) with MLOps (ML)
- Testing the feasibility of a fully open source MLOps pipeline

8.6 AutoML Robustness Tester

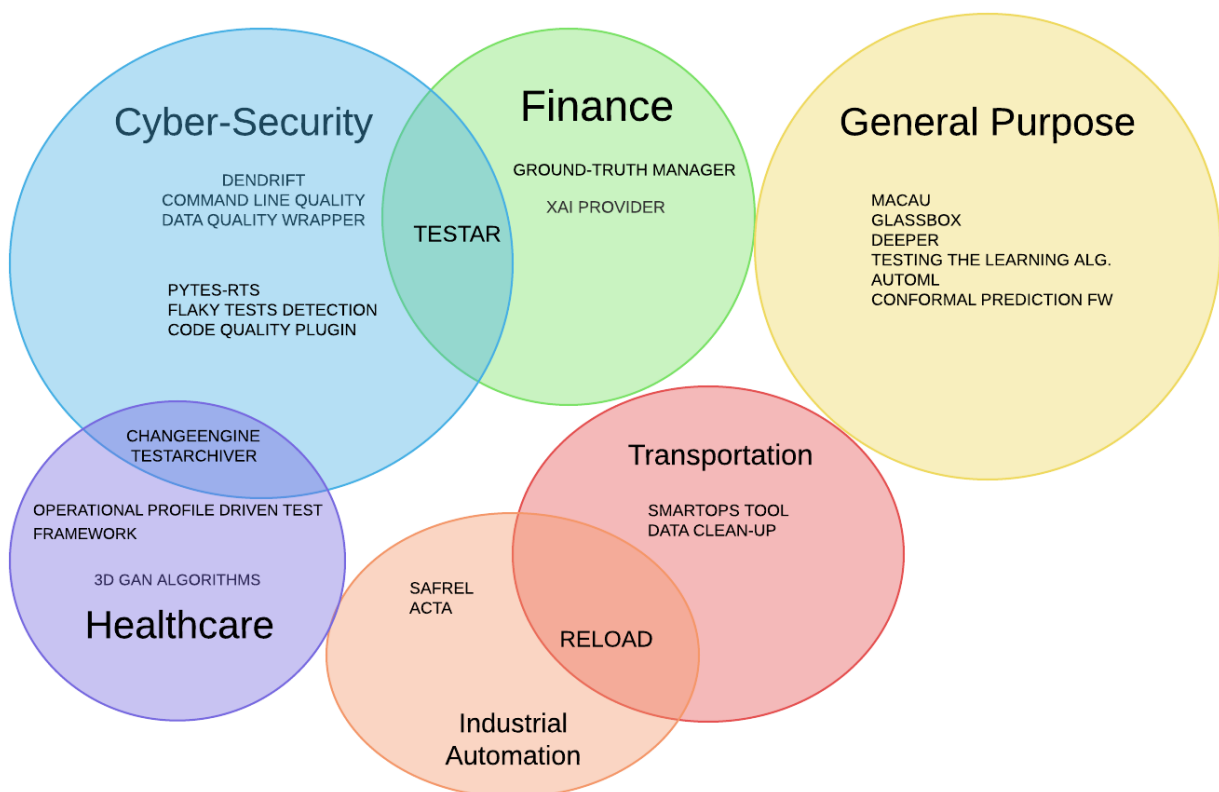A tool for testing AutoML system's robustness to different data faults.

8.7 Conformal Prediction Framework

A framework to make predictions with confidence in safety critical systems. The tool produces well calibrated error bounds, which aids towards validation of evolving systems. It is a proprietary tool, and it is opened for testing for the IVVES consortium.

9. Conclusions

In this document, a first view of the components, libraries and techniques that have been preliminary integrated as part of the validation and verification workflows has been provided. Also, the solutions that can be applied to any of the domains, have been shown. These "General Purpose" components could be eventually provided as a reference toolbox that could support new AI-based project beyond the current domains. The intuition is that, starting from a domain-specific approach, we are firstly showing a clear domain-focused V&V toolchain references, that could guide third parties to set up a specific learning (and experimentation) path. And, secondly, it is a first step that may eventually result in a set of integrated methodology, best practices and blueprints.