# Industrial Machine Learning for Enterprises

## Deliverable D2.4

## Second version of Methods and Techniques for Data Collection, Processing, and Valorisation

| | |
|---|---|
| **Project title:** | IML4E |
| **Project number:** | 20219 |
| **Call identifier:** | ITEA AI 2020 |
| **Challenge:** | Safety & Security |

| | |
|---|---|
| **Work package:** | WP2 |
| **Deliverable number:** | D2.4 |
| **Nature of deliverable:** | Report |
| **Dissemination level:** | PU |
| **Internal version number:** | 1.0 |
| **Contractual delivery date:** | 2024-05-30 |
| **Actual delivery date:** | 2024-05-30 |
| **Responsible partner:** | Software AG |

**Contributors**

| | |
|---|---|
| Editor(s) | Mohamed Abdelaal (Software AG) |
| Contributor(s) | Mohamed Abdelaal (Software AG), Timo Sinisalmi (Basware), Luca Szegletes (BUTE), Jürgen Großmann (Fraunhofer), Dorian Knoblauch (Fraunhofer), Abhishek Shrestha (Fraunhofer), Heikki Ihasalo (Granlund), Jukka Remes (Silo AI), Harry Souris (Silo AI), Kimmo Sääskilahti (Silo AI) |
| Quality assuror(s) | Lucy Ellen Lwakatare (University of Helsinki) |

**Version history**

| Version | Date | Description |
|---|---|---|
| 1.0 | 30.05.2024 | Initial full version |

**Abstract**

The deliverable discusses the evolution of the tools and technologies developed within IML4E WP2. The deliverable covers the tools which serve the objectives of the main three tasks in WP2, including data preparation automation (Task 2.1), data management and version control (Task 2.2), and continuous data quality assurance (Task 2.3).

**Keywords**

Data preparation, data valorisation, anomaly detection, data quality assurance, ML certification, quality metrics, data cleaning, data privacy

# Executive Summary

The primary objective of IML4E WP2 has been to establish a robust data layer tailored to meet the specific requirements of various use cases through the development of advanced tools and techniques. This report provides a comprehensive overview of the final state of the tools and techniques developed under WP2 as the project concludes:

- Automated Data Cleaning Tool from Software AG: This tool has been successfully implemented to automate the detection and correction of errors in tabular datasets. Utilizing state-of-the-art algorithms, it ensures high data accuracy and significantly reduces the need for manual data cleansing efforts.
- Data Quality Dashboard from Granlund Oy: This tool offers continuous monitoring of large time series data. It provides detailed analytics and real-time insights into data quality, enabling effective and proactive data management.
- Mosquito Data Cleaner from Basware Oyj: Targeted at digital invoice processing, this tool now robustly performs automated error detection and correction, ensuring the accuracy and reliability of financial data. It has become an essential component for enterprises dealing with high volumes of invoice data.
- Privacy-Friendly Human Pose Estimator from BUTE: This tool analyses human poses while strictly adhering to privacy regulations. It processes visual data effectively, maintaining anonymity and privacy, making it suitable for applications where user confidentiality is paramount.
- Continuous Audit-Based Certification (CABC) Tool from Fraunhofer FOKUS: This tool offers continuous data quality assurance. With its ability to perform continuous audits, it certifies the integrity and compliance of data throughout its lifecycle, ensuring ongoing adherence to quality and regulatory standards.

These tools collectively enhance the data layer's integrity, quality, and compliance, as envisioned in WP2. With their development complete, they are set to be integrated into broader systems, where they will continue to support and enhance data-driven decision-making processes.

# Table of contents

# Introduction

## 1.1 Role of this Document

The purpose of this document is to offer an enhanced description of the updated methodologies and techniques for data collection, processing, and valorisation within the IML4E project. It complements the detailed technical descriptions provided for each tool, which are available in the respective GitHub README files. These tools are software solutions developed as part of the project's scope. Additionally, the methodology for deploying these tools was initially outlined in Deliverable D2.2 and has been further elaborated and refined in this current document, Deliverable D2.4. The report is organized into two main sections. In the first section, we provide a structured overview of the characteristics of the WP2 tools, detailing their functionalities and intended applications. The second section delves deeper into the specifics of these tools, offering a comprehensive analysis of their technical features and operational capabilities.

## 1.2 Intended Audience

The intended audience of the present document is composed primarily of the IML4E consortium for the purpose of understanding the tools and advancing data collection, processing, and valorisation. However, this document is public and can provide an overview of the advances in the IML4E project to a wider audience. This document describes methods and technologies for the technically oriented audience rather than the general public or layman.

## 1.3 Definitions and Interpretations

The terms used in this document have the same meaning as in the contractual documents referred in [FPP] with Annexes and [PCA] unless explicitly stated otherwise.

## 1.4 Applicable Documents

| Reference | Referred document |
|---|---|
| [FPP] | IML4E – Full Project Proposal 20219 |
| [PCA] | IML4E Project Consortium Agreement |
| [D2.1] | Baseline methods and techniques for data collection, processing, and valorisation |
| [D2.2] | First version of methods and techniques for data collection, processing, and valorisation |
| [D2.4] | Second version of methods and techniques for data collection, processing, and valorisation |

**Table 1: Contractual documents.**

# 2 Characteristics of the WP2 Tools

## 2.1 Automated Data Cleaning Tool for Tabular Data

| General Information | |
|---|---|
| **Title** | Automated Data Cleaning for Tabular Data in ML Pipelines |
| **Partners** | Software AG |
| **Research area(s)** | Data Preparation |
| **Description** | In this achievement, Software AG contributed with a data cleaning framework that encompasses three different tools, called SAGED, AutoCure, and ReClean, dedicated for tabular data, such as IoT data and financial records. SAGED is a meta-learning error detection tool leveraging historical knowledge to identify and locate dataset errors. AutoCure is a data-augmentation tool enhancing data quality by increasing the density of clean instances. Finally, ReClean is a reinforcement-learning pipeline orchestration tool that optimizes data cleaning tool selection based on downstream ML model performance. |
| **Innovation** | ☒ I1: High quality and interoperable data preparation infrastructures for trustworthy ML<br>☐I2: Scalable MLOps techniques and tools for critical application domains<br>☐ I3: An MLOps Methodology<br>☐I4: An experimentation and training platform<br>☐I5: Pre-standardization work on cross-domain engineering for AI-systems |
| **Related KPIs** | ☒ ML service and process automation<br>☐Increased service delivery capability/new products<br>☐Human or/and computational resources<br>☐Effectiveness of data usage<br>☐ Finding defects |
| **Business Impact** | ☐ New AI enabled services<br>☒ Fast and efficient deployment of ML products and services<br>☒ Increased trust in AI enabled products and services<br>☐ New MLOps consulting service |
| **Impact** | The impact revolves around fully automating the data cleaning process, which is a necessary step in typical ML pipelines. |
| **Technology Environment** | The various components of the data cleaning tool have been developed using Python3 and tested on Linux machines. They have been examined with the craft beers data set, the individual income tax data set, and the smart factory data set. |
| **Synergies** | Data Quality Dashboard |
| **Access** | ☐Proprietary/Confidential ☒ Open source/access: <CC-BY 4.0> |
| **Link** | • SAGED: https://github.com/mohamedyd/SAGED<br>• AutoCure: https://github.com/mohamedyd/AutoCure<br>• ReClean: https://github.com/mohamedyd/ReClean |

| | |
|---|---|
| | • ITEA achievement: https://itea4.org/community/project/achievement/14754.html |

## 2.2 Data Quality Dashboard

| General Information | |
|---|---|
| **Title** | Data and model monitoring dashboard |
| **Partners** | Granlund, Software AG |
| **Research area(s)** | ML application monitoring and maintenance |
| **Description** | The data and model monitoring dashboard is a service that supports machine learning systems working on a large number of models. It is built on Grafana and displays crucial information about model performance, drifts, and other metrics. Data monitoring helps to understand the data and minimize the negative impact on the service. The dashboard also includes infrastructure monitoring, providing information about workflows and resources in production. It is a valuable tool for ensuring the proper function of machine learning systems. The work was aided by SoftwareAG by study of model drift method |
| **Innovation** | ☒I1: High quality and interoperable data preparation infrastructures for trustworthy ML<br>☐I2: Scalable MLOps techniques and tools for critical application domains<br>☐I3: An MLOps Methodology<br>☐I4: An experimentation and training platform<br>☐I5: Pre-standardization work on cross-domain engineering for AI-systems |
| **Related KPIs** | ☒ ML service and process automation<br>☐Increased service delivery capability/new products<br>☐Human or/and computational resources<br>☐Effectiveness of data usage<br>☒ Finding defects |
| **Business Impact** | ☐ New AI enabled services<br>☐ Fast and efficient deployment of ML products and services<br>☒ Increased trust in AI enabled products and services<br>☐ New MLOps consulting service |
| **Impact** | It helps with monitoring and fault detection of ML models, allowing for timely intervention and resolution of issues. This reduces downtime and improves customer satisfaction. Impact isn't quantifiable |
| **Technology Environment** | Grafana, EvidentlyAI, Prometheus, MLflow |
| **Synergies** | WP3 |
| **Access** | ☒ Proprietary/Confidential ☐ Open source/access |
| **Link** | |

## 2.3 Mosquito Data Cleaner

| General Information | |
|---|---|
| **Title** | Mosquito Ground Truth Refinery and Quality Feedback Service |
| **Partners** | Basware Oy, University of Helsinki |
| **Research area(s)** | Data valorisation |
| **Description** | In the scope of WP2, Basware has upgraded its Mosquito data cleaning solution from v1 to v5. The first production version v3 is based on heuristic anomaly detection algorithms featuring Two-Dimensional Context-Free Grammars for invoice composition representation. The solution was successfully deployed to production at the end of 2023, which boosted accuracy by 30% and enabled single-click learning feature for immediate error corrections. The next version is fully ML-based solution featuring self-supervised model for invoice composition representation. The prototype model Mosquito v5 has been trained, and its single-click learning coverage has been evaluated at 70%. |
| **Innovation** | ☒ I1: High quality and interoperable data preparation infrastructures for trustworthy ML<br>☒ I2: Scalable MLOps techniques and tools for critical application domains<br>☐ I3: An MLOps Methodology<br>☐ I4: An experimentation and training platform<br>☐ I5: Pre-standardization work on cross-domain engineering for AI-systems |
| **Related KPIs** | ☒ ML service and process automation<br>☐ Increased service delivery capability/new products<br>☐ Human or/and computational resources<br>☒ Effectiveness of data usage<br>☒ Finding defects |
| **Business Impact** | ☐ New AI enabled services<br>☐ Fast and efficient deployment of ML products and services<br>☒ Increased trust in AI enabled products and services<br>☐ New MLOps consulting service |
| **Impact** | Increased accuracy of supervised model for invoice data extraction due to removing human errors from training data. Instantaneous error corrections by customer request with single-click learning. |
| **Technology Environment** | ENV: Linux, TensorFlow, Python3. TECH: 2D-CFG, 2D-SCFG, Contrastive Representation Learning, Barlow Twins, BYOL, DINO, GLOM, ViT, BERT, VDBMS (ANN) |
| **Synergies** | Mosquito GT Refinery UI, Self-Validation UI |
| **Access** | ☒ Proprietary/Confidential ☐ Open source/access: |
| **Link** | Brief public technology description in “Data cleaning for supervised learning” lecture at University of Helsinki https://youtu.be/2lgn3VK1n4g<br>The official whitepaper is coming soon. |

## 2.4 Privacy-friendly Human Pose Estimator

| General Information | |
|---|---|
| **Title** | Diffusion Model-Based Facial Anonymization in TinyML setup |
| **Partners** | Budapest University of Technology and Economics (HUN) |
| **Research area(s)** | Machine learning, privacy, computer vision |
| **Description** | We've been researching the possibilities to make our initial diffusion model-based facial image anonymization method faster in terms of image generation, and less memory consuming. We decided to use a sampling technique which enables up to 50x faster image generation and significantly reduced memory consumption. Furthermore, our new approach does not compromise on image quality either. We performed thorough testing and compiled a scientific paper from our results, which we submitted to an international journal. |
| **Innovation** | ☐I1: High quality and interoperable data preparation infrastructures for trustworthy ML<br>☐I2: Scalable MLOps techniques and tools for critical application domains<br>☒ I3: An MLOps Methodology<br>☐I4: An experimentation and training platform<br>☐I5: Pre-standardization work on cross-domain engineering for AI-systems |
| **Related KPIs** | ☒ ML service and process automation<br>☐Increased service delivery capability/new products<br>☐Human or/and computational resources<br>☐Effectiveness of data usage<br>☒ Finding defects |
| **Business Impact** | ☐ New AI enabled services<br>☐ Fast and efficient deployment of ML products and services<br>☒ Increased trust in AI enabled products and services<br>☐ New MLOps consulting service |
| **Impact** | This enhancement of the diffusion model enables new, potential deployment targets among TinyML devices. |
| **Technology Environment** | Environment: UNIX-based operating system, python environment (3.11), python packages defined in the requirements.txt file in the repository.<br><br>Technologies: Diffusion models, pose estimation, deep learning |
| **Synergies** | Face anonymization in embedded environments |
| **Access** | ☐Proprietary/Confidential ☒ Open source/access: MIT licence |
| **Link** | Repository URL: https://github.com/balazsmorv/facediffusion<br>Corresponding ITEA achievement: https://itea4.org/community/project/achievement/14733.html |

## 2.5 CABC (Data Quality Attributes)

| General Information | |
|---|---|
| **Title** | Data Quality Assessment Module |
| **Partners** | Fraunhofer (DEU) |
| **Research area(s)** | Quality Assessment |
| **Description** | The Data Quality Assessment Module is a comprehensive solution for evaluating the quality of data. It measures data quality based on the stringent quality measurements specified in ISO 25024 and the latest ML-specific metrics from emerging standards. Currently, 40 computable measurements have been identified, with 4 of them specifically implemented for image-based data. Throughout the project lifetime, we are continuously working to add more measurements to the suite, ensuring that the final product stays up-to-date with the latest advancements in data quality evaluation |
| **Innovation** | ☐I1: High quality and interoperable data preparation infrastructures for trustworthy ML<br>☐I2: Scalable MLOps techniques and tools for critical application domains<br>☒ I3: An MLOps Methodology<br>☐I4: An experimentation and training platform<br>☐I5: Pre-standardization work on cross-domain engineering for AI-systems |
| **Related KPIs** | ☒ ML service and process automation<br>☐Increased service delivery capability/new products<br>☐Human or/and computational resources<br>☐Effectiveness of data usage<br>☒ Finding defects |
| **Business Impact** | ☐ New AI enabled services<br>☐ Fast and efficient deployment of ML products and services<br>☒ Increased trust in AI enabled products and services<br>☐ New MLOps consulting service |
| **Impact** | The Data Quality Assessment Module improves data quality, increases efficiency, ensures compliance with standards, and performs quality checks during the ML life cycle and development. |
| **Technology Environment** | Kubernetes, Kubeflow |
| **Synergies** | PipelineProbe |
| **Access** | ☒ Proprietary/Confidential ☐ Open source/access |
| **Link** | • https://gitlab.fokus.fraunhofer.de/ml-cse/datatessuite<br>• IML4E \| Data Quality Evaluation Tool |

# 3 Description of the WP2 Tools

## 3.1 Automated Data Cleaning Tool for Tabular Data

In today's data-driven world, enterprises and organizations across various industries broadly rely on data to drive business growth and gain a competitive edge. Data-intensive industries, e.g., banking, insurance, retail, and telecoms, typically collect diverse types of data, including sensory readings, financial records, and medical reports, to automate business tasks, facilitate better decision-making, understand performance, and satisfy customer requirements. To reap these benefits, businesses leverage analytics and business intelligence tools to extract hidden patterns or effectively predict trends and future events. However, the conclusions drawn from these tools can be misleading when the collected data contains error profiles. Real-world data often contain heterogeneous error profiles that may emerge during data collection or transfer. Some common data quality problems include missing values, duplicates, numerical outliers, inconsistencies, and violation of business and integrity rules. Consequently, ensuring the accuracy and reliability of the collected data becomes an essential prerequisite for effective data-driven applications.

During the IML4E project, we developed several innovative tools to enhance the automation of tabular data preparation. One such tool is a meta-learning-based error detection system, called SAGED[1], that automatically identifies erroneous data instances by leveraging historical knowledge. This tool significantly reduces the manual effort required to locate and correct data errors. In addition, we created a data augmentation-based data curation tool, called AutoCure, that improves the overall quality of tabular data. This tool works by augmenting the proportion of clean data instances within the dataset, effectively increasing the signal-to-noise ratio and facilitating more accurate ML models. Furthermore, we developed a cutting-edge reinforcement learning-based tool, called ReClean, that automatically selects the most appropriate data cleaning tools based on the performance of downstream ML models. This tool optimizes the data cleaning process by considering the specific requirements and characteristics of the target ML task, ensuring optimal results. In the following sections, we delve into the technical details and functionalities of each of these tools, providing a comprehensive understanding of their inner workings and potential applications in the field of data preparation and ML.

SAGED is a two-phase approach designed to identify and detect errors in tabular datasets (cf. Figure 1). The first phase, known as the knowledge extraction phase, focuses on training a series of ML models, specifically binary classifiers, for each column in the historical datasets. These models are trained to distinguish between erroneous and clean instances within the historical data, capturing the patterns and characteristics of errors specific to each column. Once the knowledge extraction phase is complete, SAGED proceeds to the detection phase. This phase begins by carefully selecting a subset of the pre-trained models that are most relevant to the input dirty dataset. The selection process involves a rigorous matching procedure that aligns the characteristics of the dirty dataset with those of the historical datasets. By choosing pre-trained models that have been trained on datasets with similar error patterns and characteristics, SAGED ensures that the selected models possess the necessary knowledge to effectively detect and address the specific errors present in the input dirty dataset.

[1] SAGED is an abbreviation for Software AG Error Detector

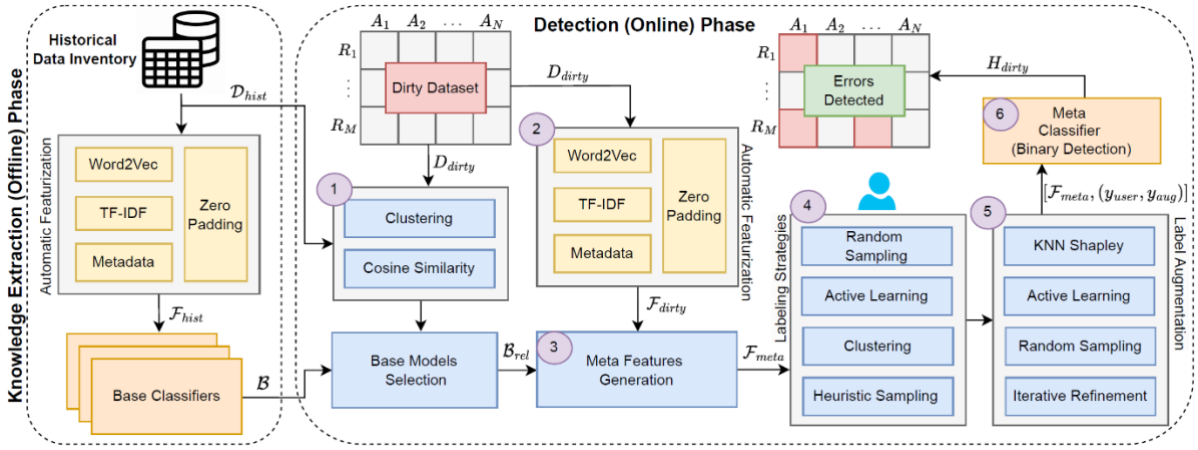


**Figure 1: Architecture of SAGED showing the knowledge extraction phase and the detection phase.**

After the appropriate pre-trained models have been selected, SAGED employs them to generate high-level feature vectors. These feature vectors are essentially the predictions generated by the selected base pre-trained models when applied to the input dirty dataset. The predictions encapsulate the insights and patterns learned from the historical datasets, serving as an abstract representation of the knowledge extracted from the base pre-trained models. By utilizing these feature vectors, SAGED captures the collective knowledge and expertise gained from the historical datasets, enabling it to identify errors in the input dirty dataset more effectively. Finally, the generated feature vectors are used to train meta-classifiers, which are responsible for precisely detecting errors in each column of the input dirty dataset. These meta-classifiers leverage the knowledge embedded in the feature vectors to make accurate predictions about the presence of errors in each instance of the dirty dataset. By combining the insights from multiple pre-trained models and utilizing the generated feature vectors, SAGED achieves a high level of accuracy in identifying and locating errors within the input dirty dataset.

Aside from SAGED, AutoCure is a data curation system that consists of two primary modules: an adaptive ensemble-based error detection module and a clean data augmentation module. The main objective of AutoCure is to synthetically increase the density of the clean fraction within the input data, which typically contains a mix of clean and dirty (erroneous) data instances. The foundation of AutoCure lies in the principles of information theory, which suggests that the process of noise reduction is equivalent to adding more data of similar quality. Building upon this theory, AutoCure aims to increase the proportion of clean data instances to mitigate the impact of noisy or erroneous data instances on the overall data quality. The adaptive ensemble-based error detection module in AutoCure is responsible for identifying and isolating the erroneous data instances within the input dataset. This module employs an ensemble of error detection algorithms that adapt to the specific characteristics and patterns of the input data. By accurately identifying the erroneous instances, AutoCure can effectively separate the clean and dirty fractions of the data.

Once the erroneous instances have been identified, AutoCure's clean data augmentation module comes into play. Instead of attempting to repair or correct the erroneous instances, which can be a complex and error-prone process, AutoCure focuses on augmenting the clean fraction of the data. By generating synthetic clean data instances that closely resemble the characteristics and distribution of the existing clean data, AutoCure effectively increases the density of clean instances within the dataset. By augmenting the clean data fraction, AutoCure reduces the relative impact of the noisy or erroneous instances on the overall data quality. This approach circumvents the challenges associated with data repair techniques and instead leverages the power of clean data augmentation to enhance the reliability and usability of the dataset.

Finally, ReClean is an innovative approach that formulates data cleaning as a sequential decision process, empowering reinforcement learning (RL) agents to select optimal repair operations based on their impact on ML model convergence. The core objective of ReClean is to maximize the predictive utility of the target application by choosing the most effective data repair tools or tool combinations. In ReClean, RL agents are trained to make sequential decisions, selecting actions that maximize the cumulative reward over an episode of cleaning and prediction steps. The reward is determined by evaluating the performance of the ML model using metrics such as AUC or R2. By optimizing these metrics, the RL agents learn to identify and apply the most beneficial data

repair techniques for the specific ML task at hand. The action space in ReClean encompasses a wide range of commonly used data repair techniques, including imputation of missing values, smoothing of outliers, encoding of textual fields, and more. These techniques are implemented as actions that the RL agents can choose from during the sequential decision process.

To learn stochastic repair policies in an off-policy manner, ReClean combines experience replay with the Reinforce algorithm. This approach allows the agents to learn from their past experiences without requiring exact environment definitions. By leveraging experience replay, the agents can efficiently explore and exploit different repair strategies, adapting to the specific error patterns present in the dataset. Through iterative interactions with the cleaned data, ReClean enables the RL agents to implicitly discover repair strategies tailored to specific error patterns without the need for direct supervision. This iterative process allows the agents to continuously refine and optimize their repair actions based on the feedback received from the ML model's performance. What sets ReClean apart from prior unidirectional methods is its joint optimization of data cleaning and ML prediction. Instead of treating data cleaning as a separate preprocessing step, ReClean integrates it into the ML pipeline, allowing the repair actions to be guided by their impact on the final predictive performance. This holistic approach ensures that the cleaned data is optimally suited for the specific ML task, leading to improved model accuracy and reliability.

## 3.2 Data Quality Dashboard

Granlund's involvement in the IML4E project centres on predicting and detecting anomalies in building energy consumption. The data quality from buildings often presents issues for machine learning models because it's sometimes stored in ways that save space at the expense of detail. Traditionally, there wasn't much need for detailed, hourly consumption data. As a result, this data can be of low quality, which might not be utilised for ML application. Additionally, it is common to have meter connection issues, and where data can be delayed, affecting our daily model operations.

To tackle these problems, we set up a central system to first check if the data from a specific building is good enough for machine learning use. Once the data is confirmed for use, the system cleans up any errors and keeps track of them. It also keeps an eye on any late-arriving data. For technology, we use custom Python scripts to manage data quality and detect any outliers, and Grafana for data visualization. The data is pulled from our databases, which collect it via APIs. We also use EvidentlyAI to detect any changes in the data trends. All these operations are run on the Microsoft Azure platform.

## 3.3 Mosquito Data Cleaner

Basware offers an invoice data extraction service SmartPDF AI to its customers. When an invoice comes to the invoice data extraction pipe, it passes up to four layers of processing. Each layer falls back to the next one if it fails to process the invoice. The first layer applies formal rules defined manually with regular expressions. It has very high accuracy, but limited coverage. If an invoice is not processed by formal rules, it falls back to ML-based processing, which consists of two layers, following two different ML paradigms: the Human-in-the-loop paradigm and the classic trainable-ML paradigm. The Human-in-the-loop model called “Mosquito” offers a highly customizable and interactive data extraction with limited coverage, while the classic trainable model called “Zebrafish” offers a generic solution that can extract data from completely unknown invoices. Finally, if both models fail, the invoice falls back to manual processing, which produces the training data for both models.

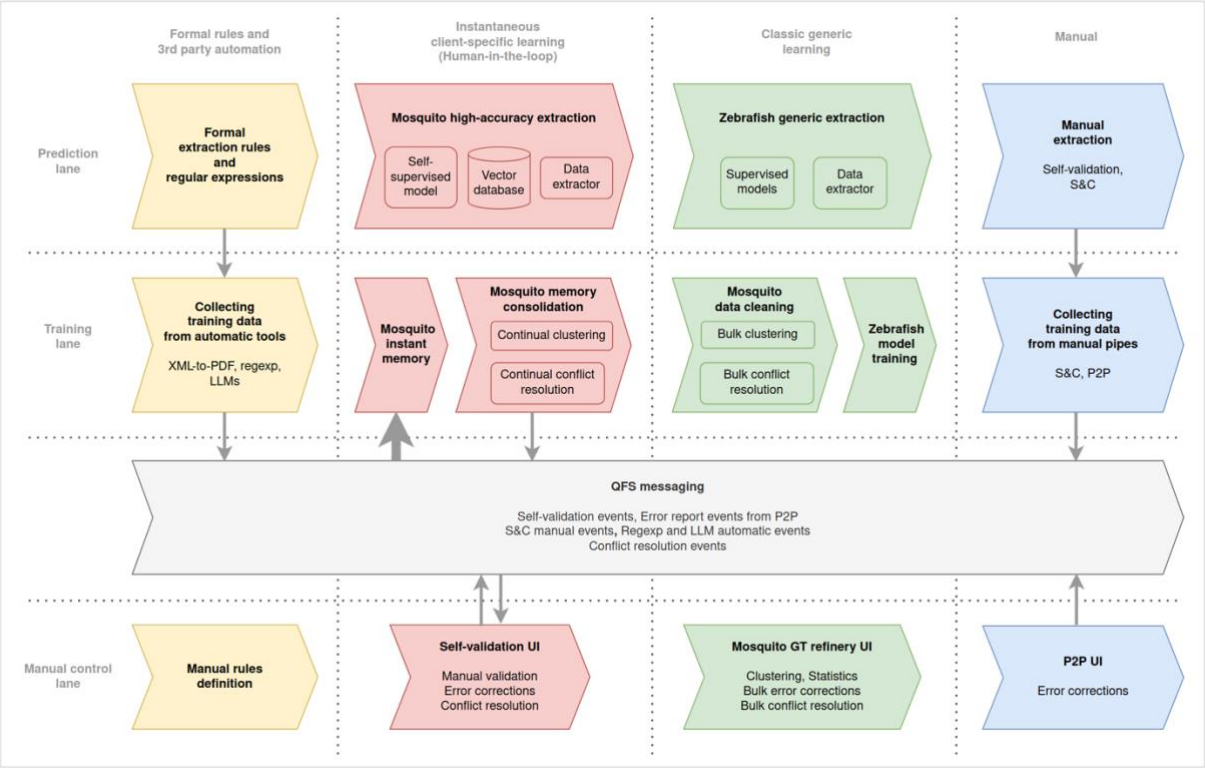


**Figure 2. Prediction, training and manual control aspects of four Basware SmartPDF AI invoice data extraction fallback layers**

The manually collected training samples contain human errors and require cleaning. The cleaning is done by clustering the invoices and detecting logical conflicts between training samples within the cluster. The conflicts can be resolved automatically with majority voting, or they can be resolved interactively with Mosquito GT Refinery UI or Self-Validation UI.

Clustering and conflict resolution is a core functionality of Mosquito model, which is used in two flavours. First is Human-in-the-loop type of data cleaning, when the error is spotted by the customer and immediately corrected. The system may detect a logical contradiction between user-defined instructions, in which case it will ask user to resolve the contradiction manually. Second is bulk offline data cleaning, which involves clustering of tens of millions of invoices and resolving the conflicts automatically with majority voting.

Both types of cleaning are done by the same core Mosquito algorithms. The interactive part is controlled by Self-validation UI, and the bulk offline part is controlled by Mosquito GT Refinery UI.

The essence of Mosquito solution is clustering based on some kind of invoice characteristics (signatures or embeddings). The level of the model that computes signatures or embeddings is called L1, the level that does clustering is called L2 and the level responsible for instant learning and conflict resolution is called L3. There are five versions of Mosquito using completely different approaches to computing invoice characteristics:

- **MG1** – bag-of-words signatures
- **MG2** – geometry-based signatures
- **MG3** – 2D-CFG grammar-based signatures
- **MG4** – grammar-based supervised embedding vectors
- **MG5** – unsupervised self-discovered embedding vectors

The versions MG2 to MG5 were implemented under WP2 package:

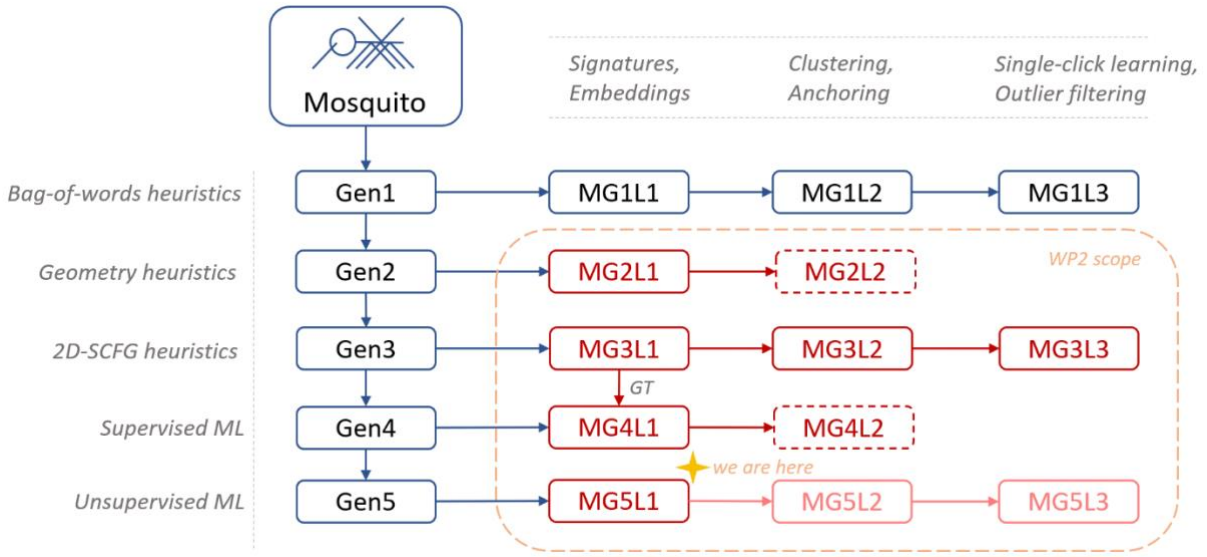


**Figure 3. Various generations of Mosquito**

The layers MG5L2 and MG5L3 are still work-in-progress. The model MG5L1 has been recently trained to the production level, and it shows ability to facilitate single-click learning with 65-70% coverage, according to preliminary evaluation. MG3 is currently in production.

## 3.4 Privacy-friendly Human Pose Estimator

In the world of big data, large amounts of data are collected, by many devices, for various applications, in almost every field in the industry. These data often contain personal information, i.e. information that can be linked to one's identity. Privacy is the fundamental human right to control how much and what information of an individual can be collected or stored. The concept first raised some interest in 1890 due to sensationalist journalism and photography. Today, privacy is arguably more important than ever, with the big data phenomenon, as a huge amount of data gets collected and stored every day. Data may be collected for personalized advertisement, or for analytics. These data could contain location information, photos and videos depicting individuals, or even private messages. There are multiple concerns with the storage and collection of these data: companies often make money selling gathered data to third parties. These third parties can then build profiles of users, possibly using multiple sources. This enables them to infer private information, not explicitly gathered: information that the user did not consent to its collection. For example, using deep learning models, the sexual orientation of a person can be correctly inferred from a single image with 81% in the case of men, and 74% in the case of women. Companies often release anonymized datasets for research purposes. For example, AOL publicly released 20 million search queries belonging to 650.000 users in 2006. The company took care of anonymization by removing the IP addresses of the queries and replacing usernames with random identifiers. However, these anonymization measures were not enough: a face was exposed for searcher no. 4417749, using only her search queries. This reveals that even if the collected information is consensual, and some (but not enough) privacy measures are taken, **data can still leak a person's identity**.

Two-dimensional human pose estimation is a challenging sub-field of computer vision. The goal is to predict a 2D skeleton for people on images and videos. The predicted skeleton consists of keypoints, which usually denote joints in the human body. These keypoints are then connected to get the person's pose. Human pose estimation is widely used in healthcare, sports, and augmented reality applications. Practical use of pose estimation however faces some privacy concerns, as the identification of facial keypoints mean that the subject's face has to be visible on the images, which can be problematic in medical settings. To solve this issue, we worked on a facial anonymization technique, which replaces subjects' face with a generated one, such that the original and synthetic facial keypoints remain on the same locations, which is a crucial aspect for pose estimation.

During the IML4e project, we developed this novel pipeline for facial image anonymization, called FaceDiffusion. One key aspect of this pipeline is the capability of generating images in a computationally constrained, TinyML environment, meaning our solution can be used on single-board computers and consumer notebooks. The pipeline is a combination of two models:

- A denoising diffusion probabilistic model (DDPM), and in a revised pipeline a denoising diffusion inference model (DDIM), capable of generating 64×64 resolution anonymized images. The model's input consists of the non-anonymized image, a bounding box of the subject's face, and facial keypoints.
- An image super-resolution model, capable of 4x upscaling and image detail restoration.

The diffusion model follows the U-Net architecture, and the implementation is based on HuggingFace's diffusion model implementation (Rogge et al., 2022). As image super-resolution model, we experimented with DFDNet (Li et al., 2020) and ESPCN (Shi et al., 2016), given the former's state-of-the-art performance, and the latter's low memory and computation consumption. The pipeline's input consists of a 64x64 resolution image to be anonymized, keypoint locations for the face, and facial bounding box. From these inputs, the model first generates an anonymised, synthetic face in low resolution, which is then upscaled by the image super-resolution model to 256x256. Some examples to the image generation can be seen in Figure 4, where the top row images are the original ones, the middle images are the low-resolution generated images, and the bottom row contains the upscaled version of the anonymized images.

**Figure 4. Comparison of various versions of facial images**

Our pipeline with the DDIM diffusion model and the DFDNet super-resolution component achieves state-of-the-art results in keypoint-location keeping ability in the root mean squared error (RMSE), the detection error (DE) and the object keypoint similarity (OKS) metrics, meaning that the original and generated images have keypoint locations very close to each other, meaning that applying pose estimation on the anonymized dataset does not come with a drawback. Table 1 contains these results, where the DDPM and DDIM image generators are our models, applied with different image super-resolution components. The results are compared with a state-of-the-art generative adversarial network, DeepPrivacy (Hukkelås et al., 2022):

**Table 1 . Evaluation results**

| Generator | Upsampler | FID | RMSE | DE | OKS |
|---|---|---|---|---|---|
| DDPM | Bicubic | 62.46 | 31.99 | 0.027 0.017 | 0.38 |
| DDPM | DFDNet | 33.19 | 17.65 | 0.0140.008 | 0.55 |
| DDPM | ESPCN | 49.2 | 47.57 | 0.0440.025 | 0.12 |
| DDPM-EMA | Bicubic | 62.45 | 30.05 | 0.0250.016 | 0.4 |
| DDPM-EMA | DFDNet | 32.91 | 17.21 | 0.0130.008 | 0.55 |
| DDPM-EMA | ESPCN | 48.08 | 46.09 | 0.0430.024 | 0.12 |

| | | | | | |
|---|---|---|---|---|---|
| DDIM-EMA | Bicubic | 57.88 | 27.29 | 0.0220.014 | 0.44 |
| DDIM-EMA | DFDNet | 33.87 | **16.75** | **0.0120.008** | **0.56** |
| DDIM-EMA | ESPCN | 42.46 | 44.66 | 0.0420.023 | 0.12 |
| DeepPrivacy | - | **13.57** | 20.37 | 0.0160.009 | 0.55 |

The metrics were calculated according to the following formulas and OKS was computed according to (COCO, 2024):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_i)^2 + \left(y_i - \overline{y}_i\right)^2}$$

$$DE = \sqrt{(x-\overline{x})^2 + (y-\overline{y})^2}/w$$

As far as image quality goes, the generative adversarial network-based solutions still outperform FaceDiffusion, which is indicated by the better Fréchet Inception Distance (FID). Our implementation is publicly available on GitHub at the following URL: https://github.com/balazsmorv/facediffusion.

## 3.5 CABC (Data Quality Attributes)

In the realm of Continuous Auditing Based Certification (CABC), measurement tools are pivotal for assessing compliance with predefined metrics critical for system certification. One such tool in this framework is the *Data Quality Evaluation Tool*, utilized to gauge the quality of data in MLOps-based AI systems. The tool is part of a broader set of tools that ensure systems align with quality dimensions such as fairness, data protection, and reliability, complying with standards like ISO, ETSI, and specific sector regulations.

The test suite includes an array of ISO metrics, notably ISO25012 and ISO25024, which focus on data quality elements such as accuracy, consistency, conformity, and completeness. For example, in the context of IoT sensor data, Deequ is used to compute accuracy measures like syntactic accuracy and semantic accuracy. The table below shows quality measures currently supported out-of-the-box for quality evaluation of tabular data.

**Table 2. Quality measures**

| Quality Characteristics | Measures | Source | Computed as |
|---|---|---|---|
| Accuracy | | ISO 25012 | X=A/B |
| | Syntactic Accuracy | ISO 25024 | A= number of syntactically correct data items<br>B= number of data items for where syntactic accuracy is required |
| | Semantic Accuracy | ISO 25024 | A= number of semantically correct data items<br>B= number of data items for where semantic accuracy is required |
| | Data Accuracy Range | ISO 25024 | A= number of data items within required range<br>B= number of data items for where accuracy in data range is required |

However, in the case of tabular data, column names, desired properties, and evaluation criteria may change overtime. For instance, in the case of Granlund, the Syntactic Accuracy is computed as the ratio of number of "DateTime" column values following YYYY-MM-DD pattern to the total rows in the dataset. Similarly, Semantic Accuracy is computed as the ratio of number of data items for which gross area is greater than the net area to the total rows. Data Accuracy Range is computed as ratio of data items for which year is between 0 and 2024 to the total number of rows. However, throughout different data versions, these requirements may change; for example the "DateTime" column may be renamed to "Date" or the YYYY-MM-DD format may change to YYYY-MM-DD hh:mm:ss while Semantic Accuracy may need to include relation between other columns or entirely

different measurement function. To provide users the flexibility of accounting for changes in datasets, the Data Quality Measurement Tool implements a rule-based quality assessments technique where the rules can be adjusted by users before each assessment process. These rules are defined in a yaml format as below:

```yaml
accuracy_syntactic:
    pattern_checks:
      - column: DateTime
        check_type: date_time_format
        parameters:
          regex: "^\\d{4}-\\d{2}-\\d{2}\\s\\d{2}:\\d{2}:\\d{2}$"
        description: "Syntactic Accuracy: (date must follow the YYYY-MM-DD format consistently)"
accuracy_semantic:
    relation_checks:
      - related_columns: ["Bruttoala", "Kerrosala"]
        check_type: relation
        parameters:
          relation: ["Bruttoala > Kerrosala"]
        description: "Semantic Accuracy: (Bruttoala > Kerrosala)"
accuracy_data_range:
    numeric_checks:
      - column: year
        check_type: range
        parameters:
          min_value: 0
          max_value: 2024
        description: "Data Accuracy range: (year < 2024)"
```

Here, the rules define which column to check for a particular measure (for example use DateTime column for syntactic accuracy) and what kind of function to use (pattern match, range checks, or column relation checks). The application adjusts dynamically to the defined rules and computes these measures accordingly. This workflow also provides users the flexibility to define their own measures and their measurement method besides the pre-defined ones.

For image datasets, tools like FiftyOne assess data quality through measures such as Intersection Over Unions (IOUs) and image hashing to detect annotation overlaps and duplicate identification, respectively. The application currently covers the following measures:

| Quality Characteristics | Measures | Source | Computed as |
|---|---|---|---|
| Accuracy | | ISO 25012 | For X = A/B |
| | Syntactic Accuracy | ISO 25024 | A= Number of images with duplicate object labels<br>B= Total number of images in the dataset |
| | Semantic Accuracy | ISO 25024 | A= Number of images with incorrect object labels<br>B= Total number of images in the dataset |
| | Data Accuracy Assurance (Syntactic) | ISO 25024 | A= Number of images measured for syntactic accuracy.<br>B= Total number of images in the dataset |
| | Data Accuracy Assurance (Semantic) | ISO 25024 | A= Number of images measured for semantic accuracy.<br>B= Total number of images in the dataset |
| Completeness | | ISO 25012 | |
| | Data file Completeness | ISO 25024 | A= Number of images in the dataset<br>B= Total number of images expected |
| | Attribute Completeness | ISO 25024 | A= Number of records with no missing annotations<br>B= Total number of images in the dataset |
| | Metadata Completeness | ISO 25024 | A= Number of records with no missing meta-data like image width, height, filename, and date captured<br>B= Total number of images in the dataset |

| | | | |
|---|---|---|---|
| Consistency | | ISO 25012 | |
| | Risk of inconsistency | ISO 25024 | A= Number of duplicate images<br>B= Total number of images in the dataset |
| | Data values consistency coverage | ISO 25024 | A= Number of images checked for consistency<br>B= Total number of images in the dataset |
| | Near Duplicates | | A= Number of images with near duplicates<br>B= Total number of images in the dataset |
| Compliance | | ISO 25012 | |
| | Regulatory Compliance of value and/or Format | ISO 25024 | A= Number of images which follow regulatory compliance (with license)<br>B= Total number of images in the dataset |

In the table above, metadata requirements, the number of expected data points, and regulatory requirements are provided as user input through a YAML file or when executing the tool. Metadata completeness, data file completeness, and regulatory compliance are computed based on these inputs. Near duplicates are identified using the open-source Python API called imagededup (https://github.com/idealo/imagededup). The library uses an imagenet trained MobilenetV3 model sliced at the last convolutional layer to compute embeddings. These embeddings or feature vectors are compared using cosine similarity (default 85%) to compute near dups.

These tools have been applied in practical scenarios such as the Grandlund use case, where the focus was on enhancing the reliability and efficiency of IoT sensor data. Similarly, in the Siemens use case, image data was scrutinized to detect semantic and syntactic labeling errors effectively using the developed tools. Such applications demonstrate the tools' capabilities in providing detailed data quality assessments, which are crucial for maintaining the certification status of AI systems in diverse operational environments. Through these measures, the tools help in systematically enforcing high data quality standards, thereby supporting better decision-making and enhancing operational efficiencies in MLOps contexts.

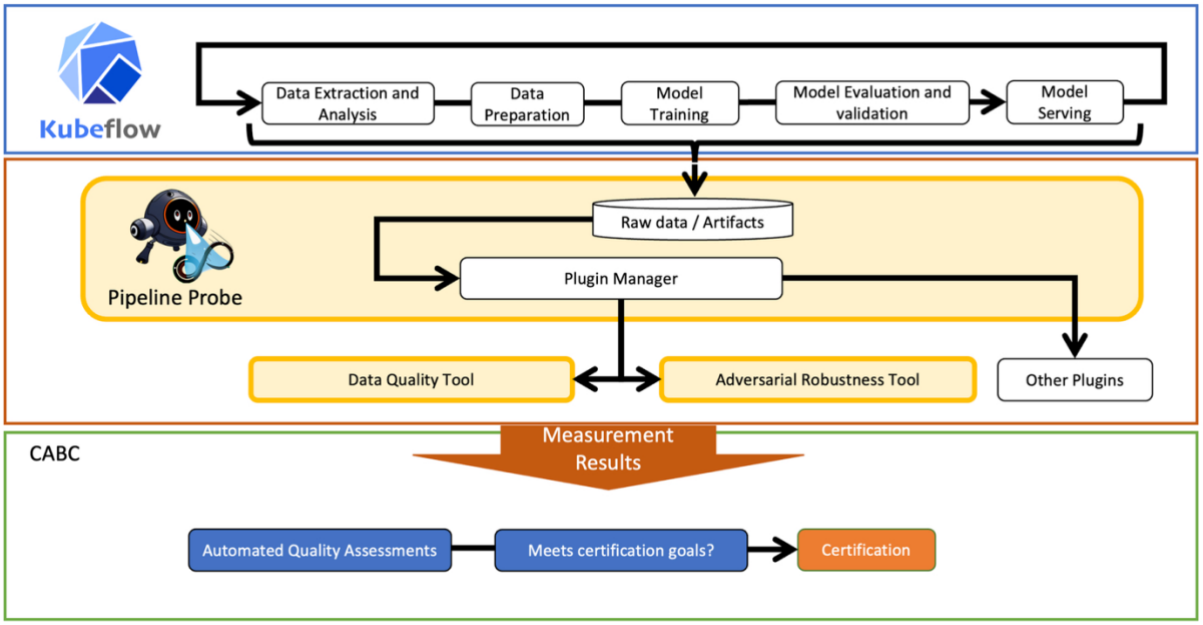


**Figure 5. data flow in an MLOps pipeline using Kubeflow**

The image depicts a data flow in an MLOps pipeline using Kubeflow, where data is processed through various stages, from extraction to serving. A key component, the Pipeline Probe, uses plugins like the Data Quality Tool to evaluate data quality by examining raw data or artifacts. The results from these assessments determine if the

system meets the required standards for certification. This setup ensures that the data and models are continuously checked for quality and compliance as part of the certification process.

# 4 Summary

This deliverable reports on the advancements and final state of the tools and techniques developed within WP2 of the IML4E project. The primary objective of WP2 has been to establish a robust data layer tailored to meet the specific requirements of various use cases through the development of advanced data preparation, management, and quality assurance tools. The key contributions and innovations presented in this deliverable include:

- Automated Data Cleaning Tool from Software AG: This tool automates the detection and correction of errors in tabular datasets using state-of-the-art algorithms. It significantly reduces manual data cleansing efforts and enhances data accuracy through meta-learning, data augmentation, and reinforcement learning techniques.
- Data Quality Dashboard from Granlund Oy: A comprehensive service built on Grafana for continuous monitoring and real-time insights into data quality. This tool aids in the proactive management of large time series data and ensures effective data monitoring and fault detection.
- Mosquito Data Cleaner from Basware Oyj: An advanced tool for digital invoice processing that performs automated error detection and correction. It employs heuristic and ML-based anomaly detection algorithms to ensure the accuracy and reliability of financial data.
- Privacy-Friendly Human Pose Estimator from BUTE: A tool that adheres to privacy regulations while analysing human poses. It uses a novel pipeline for facial image anonymization that can operate in computationally constrained environments, ensuring user confidentiality without compromising on image quality.
- Continuous Audit-Based Certification (CABC) Tool from Fraunhofer FOKUS: This tool offers continuous data quality assurance through automated audits. It certifies the integrity and compliance of data throughout its lifecycle, ensuring adherence to quality and regulatory standards.

The deliverable also provides detailed descriptions of the technical features, operational capabilities, and business impacts of each tool. These tools collectively enhance the data layer's integrity, quality, and compliance, supporting data-driven decision-making processes in various industrial applications. The methodologies and techniques outlined in this document reflect the collaborative efforts of the IML4E consortium to advance the state of data collection, processing, and valorisation in industrial machine learning applications.

# References

Niels Rogge and Kashif Rasul. The annotated diffusion model. June 2021. https://huggingface.co/blog/annotated-diffusion, Last accessed: 2024.05.02.

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1874–1883, 2016

Xiaoming Li, Chaofeng Chen, Shangchen Zhou, Xianhui Lin, Wangmeng Zuo, and Lei Zhang. Blind face restoration via deep multi-scale component dictionaries. In ECCV, 2020.

Hukkelås, H., Mester, R. and Lindseth, F., 2022. *DeepPrivacy: A Generative Adversarial Network for Face Anonymization*.

COCO dataset, Keypoint evaluation, [Online]. Available: https://cocodataset.org/#keypoints-eval, Last accessed: 2024.05.02.