

CERTAIN

Certification for Ethical and Regulatory Transparency in Artificial Intelligence

D5.1: TEST DATA GENERATION METHODS

Co-funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

Work package	WP 5
Task	Task 5.1
Due date	31/08/2026
Submission date	28/08/2026
Deliverable lead	UL
Version	1.0
Authors	Ali Kordia (UL), Tania Carvalho (UL), Izidor Mlakar (UM), Urška Smrke (UM), Valentino Šafran (UM), Fabian Kovac (SPU), Dimitrios Christodoulou (D4P), Zouhair Haddi (NVIS), Carlo Mazzola (NVIS), Ioannis Krasonikolakis (NBG), Ünsal Öztürk (IDIAP), Victoria Sovatzoglou (INCOM), Irina Urumova (INCOM), Eleftheria Tsourlidaki (NVCR), Pierre Gacon (IPS)
Reviewers	Konstantinos Amplianitis (UCD), Irina Urumova (INCOM), Victoria Sovatzoglou (INCOM), Eleftheria Tsourlidaki (NVCR)
Security reviewer	Peter Kieseberg (SPU)
Abstract	Design and implementation of advanced synthetic data generation methods to support robust compliance evaluation of AI systems under EU regulatory frameworks. The work establishes a principled approach combining generative modelling techniques with domain-specific constraints to produce realistic and representative test datasets across multiple domains, including energy, healthcare, finance, biometrics, and text. Emphasis is placed on ensuring data fidelity and diversity to effectively identify potential system vulnerabilities. The report also examines the regulatory and legal framework governing synthetic data, ensuring alignment with applicable EU requirements. The outcomes include validated algorithms and a structured methodology, delivered as part of an integrated toolbox to assist AI system providers in conducting reliable and scalable compliance assessments.
Keywords	Synthetic Test Data, AI Compliance Assessment, EU Regulatory Framework, Robustness Evaluation, Data Governance, AI System Testing

Draft

Document Revision History

Version	Date	Description of change	List of contributor(s)
V0.1	27/03/2026	1st draft	Ali Kordia (*UL*)
V1.0	28/08/2026	Completed and reviewed deliverable	Partners (*T5.1*)

© 2024-2027 CERTAIN

Grant Agreement No: 101189650 | **Topic:** HORIZON-CL4-2024-DATA-01-01
Call: HORIZON-CL4-2024-DATA-01. | **Type of action:** HORIZON-IA

REVIEWERS COMMENTS

<the comments must be addressed when the document is delivered to the EC>

Victoria Sovatzoglou	Affiliation: INCOM	Date of review: 19/06/2026 (1st round) Date of review: 14/07/2026 (2nd round)

1st review – 19.06.2026

The Deliverable is well structured providing a clear overview of synthetic data generation methods across different domains. It shows well how synthetic data can support AI compliance testing under EU regulations.

The combination of technical approaches with regulatory considerations (AI Act, GDPR) is considered as its strong point. The document also clearly links use cases, risks, and mitigation measures increasing its practical value.

However, there are few suggestions for further improvements:

- **Clarity and consistency of writing**: Some sections would benefit from linguistic refinement, clearer phrasing, and more consistent use of style (e.g., avoiding first-person narration, improving sentence structure, and standardizing terminology).
- **Formalisation of objectives and contributions:** The introduction could more explicitly state the main objectives and expected outcomes of the deliverable in a concise and structured way.

2nd review – 14.07.2026

After reviewing the latest version of D5.1, including the newly added partner contributions, the overall assessment is that the deliverable is substantially improved. The document now follows a more consistent structure across domains, moving from motivation and requirements to methodology, compliance testing, and FAIR principles. However, there are still some observations for improvements:

- The level of detail varies significantly between domains. For instance, Biometrics and Energy are extremely detailed, while some other sections are more concise.
- The Deliverable needs to be fully unified. A cross-domain synthesis in the conclusion could support this.
- Section 4: It is recommended a short introductory paragraph at the beginning of Section 4 explaining how the requirements identified in Section 3 are addressed by the methodologies presented in Section 4.
- Some glossary and acronym issues remain unresolved. Cross check needs to be done among the whole text and the table with Acronyms as some of them might be missing to be indicated, such as LDM, ID, DCFace etc.

N.B As INCOM has contributed to Section 2 (Regulatory and Legal Framework for Synthetic Data), as well as to Sections 3.6 and 4.6 corresponding to Human Resources domain, these sections were not subject to review.

Draft

© 2024-2027 CERTAIN

Konstantinos Amplianitis	Affiliation: UCD	Date of review: 19/06/2026 (1st round) Date of review: 14/07/2026 (2nd round)

1st Review – 19.06.2026

The deliverable is comprehensive and addresses an important aspect of the CERTAIN framework by defining synthetic test data generation approaches across multiple domains. The integration of technical methodologies with regulatory and compliance considerations is a particular strength. The document clearly discusses motivations, risks and mitigation measures associated with synthetic data generation and provides useful domain-specific insights.

Further harmonisation of style, terminology and structure across partner-contributed sections would improve readability and consistency. The deliverable would also benefit from a clearer statement of objectives and contributions, as well as a more explicit description of synthetic data validation and quality assessment methodologies. Specifically, Section 1.2 could be strengthened by explicitly distinguishing between the task objectives and the concrete contributions of the deliverable. For example, the introduction should clearly state what D5.1 delivers to the project, such as the domain-specific synthetic data generation approaches, the mapping of generated test data to compliance needs, the associated risks and mitigation measures, and the expected role of these outputs in supporting CERTAIN's compliance assessment toolbox.

Subject to the above revisions, the deliverable represents a valuable contribution to the CERTAIN project and provides a solid basis for compliance-oriented AI testing using synthetic data.

2nd Review – 14.07.2026

Performing a second review of this deliverable, an additional issue concerns the consistency and accuracy of the references and in-text citations. A systematic cross-check indicates that several citation numbers do not correspond to the relevant sources in the reference list. For example, on page 21, Goodfellow et al. (2014) is cited as reference [1], although reference [1] concerns the LPIPS perceptual metric, while the Goodfellow et al. publication is listed as reference [26]. Similarly, the Abay et al. publication is cited as reference [2], although it appears as reference [34]. On page 26, the discussion of the NBG/FAR-Trans dataset cites reference [8], but reference [8] concerns radar-based vital-sign monitoring, whereas the FAR-Trans publication is listed as reference [38]. All in-text citations should therefore be systematically cross-checked against the bibliography, corrected, and manually verified before submission.

The overall structure of the deliverable is logical, progressing from the regulatory and legal framework in Section 2, to domain-specific motivations, requirements, and risks in Section 3, and then to the corresponding generation methodologies in Section 4. However, the connections between these sections should be made more explicit. For each domain, the requirements and risks identified in Section 3 should be clearly traced to the generation method, validation metrics, achieved results, and compliance evidence presented in Section 4. At present, this relationship is uneven: some domains provide detailed technical and experimental evidence, particularly biometrics and energy, while others mainly describe proposed approaches, planned validation, or future releases. The deliverable should therefore distinguish consistently between implemented, validated, ongoing, and planned work.

A few cross-section issues should also be clarified. In the finance section, the relationship between the six-month ROI horizon and the future-return target calculated over 21 trading days should be explained more explicitly, as these appear to describe different elements of the methodology but may be interpreted by the reader as competing prediction horizons. The text contribution states that no real-world discussion dataset is used, whereas the conclusion indicates that each partner started from a real or reference dataset; this should be reformulated to reflect the different starting points across domains. In addition, the conclusion should avoid making claims that are stronger than the evidence reported in the technical sections, particularly regarding biometric anonymity and the existence of a fully common validation methodology.

N.B As UCD has contributed to Section 2 (Regulatory and Legal Framework for Synthetic Data) as well as to Sections 3.6 and 4.6 corresponding to Human Resources domain, together with INCOM, these sections were not subject to review.

Draft

Name of reviewer: Peter Kieseberg	Affiliation: SPU	Date of review: 3.8.2026

The introduction is very clear and well-motivated.

Please check the glossary, some terms are duplicated, like Pseudonymization (3 times), Diffusion model (3 times) ...

Chapter 2 I would expect an overview on all regulations and frameworks deemed relevant in the context of this work.

For chapter 3.1 I would expect far more references to the academic literature. The use cases are arbitrary, as well as the risks and both need motivation for their selection, maybe through a ranking in an academic reference. Also, the connection between synthetic data and regulatory requirements needs to be elaborated further. It is clear that synthetic data can help companies that are not banks and don't have access to real data, but the banks in my opinion could already have these models. Why is it beneficial to the public to have more players in this field? Is this not also a problem regarding fraud?

For 3.2, I understand that the selected use-case is highly specific, but I would have pointed out that with P4 medicine and related approaches there is an extremely high need for good models for all types of people and real data is a major bottleneck for realizing those ambitious plans.

Similar issues hold true for the remaining section. For the tables in Section 3, I am rather unclear regarding the selection of Use cases and Risks. Was this done using a literature survey? Especially regarding the risks, I, as a Security person, miss malevolent behaviour from both, insiders and external persons. Introducing Backdoors/biases into models is a very good attack vector, but besides Safety, Security as not considered in this risk analysis.

I am also unsure, why Sections 3 and 4 were not condensed into one section and maybe broken apart on application domain level.

Finally, I miss references to the academic literature throughout many parts of the work, where facts are stated but not referenced.

Name of reviewer: Eleftheria Tsourlidaki	Affiliation: NVCR	Date of review: 20.07.2026

- Some parts need some ironing out as the whole deliverable feels a bit like a patchwork rather than a unified document.
- I propose we reconsider chapter 3 and 4. The way they read now is a bit tiring, as the reader gets to read the motivation and requirements for 6 different cases and once that is done, another six-stops circle begins to read about the methodology. I would propose we merge these sections for each case and keep on chapter per case each with two main subsection; motivation and requirements, methodology and datasets delivery.
- The existing motivation and requirements I would reshape it to introduce a general section about why we need synthetic data sets in general, and then how these general needs manifest in CERTAIN within the pilot specific contexts. (This is different to chapter 2 that mentions the limitations). I would also use that section to build a more solid bridge between T5.1 and this deliverable to the pilots, taking the opportunity to explain how each pilot worked separately to produce synthetic data for their needs. This would also explain/better justify the different writing styles of following up sections. Note that there is no section at the moment clearly explaining the link between the pilots and this deliverable. While this information is apparent to us and inferred in some parts of the text, we should keep in mind that the reviewers are external to the project and we should not take any information for granted assuming they already know how we organized our work.
- I would also propose that an overview section is added (maybe a table) summarizing what was produced in total (what types of datasets, how many of each kind, main intended use, core methodology or component used) etc.

Draft

- I'm also missing the transferability dimension. While we clearly produced methodologies and datasets for the needs of our pilots, what is their added value beyond the project? Can they be used by other actors and for what reason? Is what has been produced somehow useful beyond the project's context?

Draft

DISCLAIMER



Co-funded by the European Union

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

The CERTAIN project received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement No 101189650. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. This work has received funding from the Swiss State Secretariat for Education, Research and Innovation (SERI).

COPYRIGHT NOTICE

© 2025 – 2027 CERTAIN

Project funded by the European Commission in the Horizon Europe Programme		
Nature of the deliverable:	R (report)	
Dissemination Level		
PU	Public, fully open, e.g. web (Deliverables flagged as public will be automatically published in CORDIS project's page)	☒
SEN	Sensitive, limited under the conditions of the Grant Agreement	☐

*

Draft

TABLE OF CONTENT

Reviewers comments 3
DISCLAIMER 7
Copyright notice 7
1 INTRODUCTION 19
1.1 Background 20
1.2 structure of the deliverable 20
2 REGULATORY AND LEGAL FRAMEWORK FOR SYNTHETIC DATA 21
2.1 Key terms 21
2.2 Legal status of synthetic data 21
2.3 Interaction with the AI Act 22
2.4 Residual risks 22
2.5 Core ethical principles 22
2.6 Application to the human resources pilot 23
3 MOTIVATION AND REQUIREMENTS 26
3.1 Finance 26
3.2 Health 28
3.3 Biometrics 32
3.4 Energy 33
3.5 Text 34
3.6 Human resources 35
4 GENERAL METHODOLOGY FOR SYNTHETIC TEST DATA GENERATION 38
4.1 Finance 38
4.2 Health 40
4.3 Biometrics 49
4.4 Energy 66
4.5 Text - Public Deliberation Synthetic Datasets Generation Component 88
4.6 Human resources 91
5 CONCLUSIONS 94
6 REFERENCES 95

Draft

© 2024-2027 CERTAIN

EXECUTIVE SUMMARY

AI systems now make consequential decisions in healthcare, finance, Human Resources, biometrics, and energy. Showing that these systems meet EU expectations on fairness, robustness, transparency, and data protection requires test evidence, and test evidence requires data. The data that would make such testing most informative is often the data that cannot be shared: patient physiology, customer transactions, biometric captures, household consumption traces. Real data of this kind is locked behind GDPR compliance and behind the legitimate refusal of individuals to expose their lives for the sake of model evaluation.

Synthetic data offers a way around this. Artificial datasets that reproduce the statistical and structural properties of real data, without reproducing the people behind it, can support compliance testing under conditions that resemble deployment without the privacy cost. This document reports how CERTAIN partners built such datasets across five application domains, with a sixth, human resources, providing the regulatory and theoretical foundation that grounds the work.

Draft

© 2024-2027 CERTAIN

LIST OF FIGURES

FIGURE 4-1: THE CERTAIN SYNTHETIC DATA GENERATION & EVALUATION PIPELINE. NOTE: THIS IMAGE WAS GENERATED WITH THE USE OF CHAT GPT 5.4 44

FIGURE 4-2 DET CURVES COMPARING THE PERFORMANCE OF ARCFACE AND IPS_2025 FOR A DATASET OF REAL IMAGES. 52

FIGURE 4-3 EXAMPLES OF SYNTHETIC IMAGES GENERATED BY OUR 128X128-RESOLUTION STYLEGAN2 (COLUMN 1), BY IPS2FACE-REF (COLUMN 2), AND BY IPS2FACE (COLUMNS 3, 4, AND 5). 53

FIGURE 4-4 MAGFACE MATCHING STATISTICS FOR SYNTHETIC DATASETS GENERATED USING VARIANTS OF ARC2FACE AND IPS2FACE..... 54

FIGURE 4-5 DISTRIBUTIONS OF LPIPS DISTANCES FOR THREE SYNTHETIC DATASETS REVEALING THE OVERALL PERCEPTUAL VARIETY WITHIN MATED SETS OF IMAGES. 54

FIGURE 4-6 MAGFACE MATCHING STATISTICS FOR SYNTHETIC DATASETS ARC2FACE-LIVE+REF, IPS2FACE-LIVE+REF AND FLUXSYNID..... 55

FIGURE 4-7 FALSE ACCEPTANCE RATES FOR SYNTHETIC-TO-REAL IMPOSTOR PAIRS AND REAL-TO-REAL IMPOSTOR PAIRS. SYNTHETIC IMAGES WERE GENERATED USING STYLEGAN2 + ARC2FACE (LEFT) AND STYLEGAN2 + ARC2FACE-REF (RIGHT). THE FR ALGORITHM IPS_2025 WAS USED FOR CODING AND MATCHING. 56

FIGURE 4-8 PERFORMANCE STATISTICS FOR VARIOUS IPS FR ALGORITHMS EVALUATED USING AN OPERATIONAL DATASET OF REAL IMAGES (TOP) AND OUR SYNTHETIC ARC2FACE-LIVE+REF DATASET (BOTTOM)..... 56

FIGURE 4-9 PERFORMANCE STATISTICS FOR TWO IPS FR ALGORITHMS EVALUATED ON SUBSETS OF VARYING SKINTONE FOR BOTH AN OPERATIONAL DATASET OF REAL IMAGES (LEFT) AND OUR SYNTHETIC ARC2FACE-LIVE+REF DATASET (RIGHT). 57

FIGURE 4-10 SOFT-PARTICLE CONTACT-FORCE MODEL UNDERLYING THE DISCO LANGEVIN AND DISPERSION DYNAMICS, REPRODUCED FROM THE PAPER “SYNTHETIC FACE DATASETS GENERATION VIA LATENT SPACE EXPLORATION FROM BROWNIAN IDENTITY DIFFUSION” 59

FIGURE 4-11 BLOCK DIAGRAM OF THE HYPERFACE DATASET-GENERATION PIPELINE, REPRODUCED FROM THE PAPER “HYPERFACE: GENERATING SYNTHETIC FACE RECOGNITION DATASETS BY EXPLORING FACE EMBEDDING HYPERSPHERE” [16]. 60

FIGURE 4-12 ENERGY DATA GENERATION ARCHITECTURE 75

FIGURE 4-13 ENERGY DATA MODEL SHOWING THE ENTITIES (ENERGY_COMMUNITY, PARTICIPANT, METERING_POINT, ENERGY_DATA), THEIR RELATIONS, AND THE MEASUREMENT-TYPE, ENERGY-SOURCE, AND ENERGY-DIRECTION ENUMERATIONS. 82

FIGURE 4-14 REFERENCE DATA MODEL SHOWING THE WEATHER OBSERVATIONS (GEOSPHERE INCA) AND THE EPEX SPOT 1-DAY-AHEAD AUSTRIAN MARKET PRICE, TOGETHER WITH THE WEATHER OBSERVATION-TYPE ENUMERATION..... 85

Draft

LIST OF TABLES

TABLE 1-1: GLOSSARY 14

TABLE 1-2: ACRONYMS 18

TABLE 3-1 SYNTHETIC DATA USE CASES AND THE COMPLIANCE REQUIREMENTS ADDRESSED 27

TABLE 3-2 RISKS OF SYNTHETIC DATA GENERATION AND MITIGATION MEASURES 27

TABLE 3-3 HEALTH, DIGITAL PHENOTYPING: USE CASES AND COMPLIANCE REQUIREMENTS 29

TABLE 3-4 HEALTH, DIGITAL PHENOTYPING: RISKS AND MITIGATION MEASURES 29

TABLE 3-5 HEALTH, RADAR MONITORING: USE CASES AND COMPLIANCE REQUIREMENTS 31

TABLE 3-6 HEALTH, RADAR MONITORING: RISKS AND MITIGATION MEASURES 31

TABLE 3-7 AN ASSESSMENT OF POTENTIAL SOLUTIONS FOR EVALUATION DATASET FORMATION. 32

TABLE 3-8 COMPLIANCE MAPPING 36

TABLE 3-9 RISKS AND MITIGATION MEASURES 37

TABLE 4-1: RELEVANT DATA POINTS IN THE COHORTS TO BE CONSIDERED FOR THE SYNTHETIC DATA GENERATION 42

TABLE 4-2 DATASETS OF BONA FIDE FACE IMAGES USED FOR TRAINING OUR GENERATIVE MODELS. 50

TABLE 4-3 SYNTHETIC IDENTITY COUNTS PER DEMOGRAPHIC GROUP. OUR SKINTONE SCALE SPANS FROM 0.0 (LIGHT SKIN) TO 1.0 (DARK SKIN). 51

TABLE 4-4 DATASETS AS PART OF THE TRAINING CORPORA FOR SYNTHETIC DATA GENERATION METHODS CONSIDERED BY IDIAP. 57

TABLE 4-5 DATASETS AS PART OF THE EVALUATION OF SYNTHETIC DATA GENERATION METHODS CONSIDERED BY IDIAP. 58

TABLE 4-6 QUALLITATIVE COMPARISON OF STATE-OF-THE-ART METHODS TO DISCO AND HYPERFACE. 61

TABLE 4-7 DOWNSTREAM FACE-RECOGNITION ACCURACY FOR EACH METHOD AT THE SAME SCALE OF SYNTHETICALLY GENERATED DATASETS. THE LAST ENTRY ALLOWS THE COMPARISON OF THE PERFORMANCE OF SYNTHETIC DATASETS WITH REAL DATASETS. 62

TABLE 4-8 DOWNSTREAM FACE-RECOGNITION ACCURACY FOR EACH METHOD AT DIFFERENT SCALES OF SYNTHETICALLY GENERATED DATASETS. THE LAST TWO ENTRIES ALLOW THE COMPARISON OF THE PERFORMANCE OF SYNTHETIC DATASETS WITH REAL DATASETS. 63

TABLE 4-9 DEMOGRAPHIC BIAS DUE TO ETHNICITY ON THE RFW BENCHMARK WITH OTHER STATE-OF-THE-ART METHODS THAT REPORT A BENCHMARK ON RFW. 63

TABLE 4-10 SUGGESTED DATA-GENERATION PROTOCOLS FOR FACE-RECOGNITION EVALUATION. DEMOGRAPHIC-FAIRNESS AXES USE IDENTITY-DISJOINT GROUPS, WITH VERIFICATION ACCURACY COMPUTED PER GROUP AND THE DISPARITY ACROSS GROUPS AS THE METRIC. COVARIATE-ROBUSTNESS AXES USE A SINGLE IDENTITY SET,

Draft

WITH VERIFICATION ACCURACY COMPUTED ACROSS IMAGE PAIRS OF THE SAME IDENTITY BUT WITH DIFFERENT COVARIATE VALUES. 64
TABLE 4-11 POSITIONING OF THE TWO SYNTHESIS REGIMES 67
TABLE 4-12 IDENTIFICATION 79
TABLE 4-13 PROVENANCE AND SOURCE FOUNDATION 80
TABLE 4-14 PUBLISHED DATASET SCOPE AND STRUCTURE 80
TABLE 4-15 PUBLISHED DATASET METADATA: EXTERNAL REFERENCE DATA SOURCES 81
TABLE 4-16 LEGAL BASIS AND RELEASE CONDITIONS 81
TABLE 4-17 ENERGY DATA SCHEMA: ENERGY_COMMUNITY ENTITY 82
TABLE 4-18 ENERGY DATA SCHEMA: PARTICIPANT ENTITY 82
TABLE 4-19 ENERGY DATA SCHEMA: METERING_POINT ENTITY 83
TABLE 4-20 ENERGY DATA SCHEMA: ENERGY_DATA ENTITY 83
TABLE 4-21 ENERGY DATA SCHEMA: PARTICIPANT_METERING_POINT RELATION 83
TABLE 4-22 ENERGY DATA SCHEMA: ENERGY_COMMUNITY_PARTICIPANT RELATION 84
TABLE 4-23 ENERGY DATA SCHEMA: PARTICIPATION_FACTOR_HISTORY RELATION 84
TABLE 4-24 ENERGY DATA SCHEMA: MEASUREMENT TYPE ENUMERATION (OBIS CODES) 84
TABLE 4-25 REFERENCE DATA SCHEMA: EPEX SPOT 1-DAY AHEAD PRICE AT 85
TABLE 4-26 REFERENCE DATA SCHEMA: WEATHER DATA (GEOSPHERE AUSTRIA INCA) 85
TABLE 4-27 REFERENCE DATA SCHEMA: WEATHER OBSERVATION TYPE ENUMERATION 86
TABLE 4-28 FEATURE CATALOGUE OF THE PLANNED OUT-OF-DISTRIBUTION GENERATIVE ENERGY DATASET 86

Draft

GLOSSARY AND ACRONYMS

The list of terms, group of words, and expressions used in this document are explained in Table 1-1 below

Term, expression	Definition
Anonymisation	Processing that irreversibly removes the link between data and an identifiable natural person, so that the data falls outside the scope of the GDPR (Recital 26).
Clean/dirty dataset pair	Two variants of the same synthetic dataset: a clean version with complete, accurate records, and a dirty version with controlled degradation (shuffled, nullified or truncated fields) used for robustness testing.
Compliance testing	Structured evaluation of an AI system against regulatory requirements such as accuracy, robustness, fairness, transparency and data protection.
Conditional generation	Synthetic data generation steered by contextual inputs (for example age group, weather, or household metadata) so that outputs satisfy predefined characteristics.
Counterfactual explanation	A minimally modified version of a real input that changes a model's prediction, used to explain what would need to change for a different outcome to be produced.
Diffusion model	A generative model that produces data by iteratively denoising random noise, optionally conditioned on context variables.
Digital phenotyping	The quantification of observable behavioural and physiological cues (speech, facial affect, gaze, physiology, environment) as measurable features associated with health-related outcomes.
Energy community (EEG)	A legal entity under the Austrian Renewable Expansion Act in which households, businesses and producers jointly generate and consume renewable electricity.
Federated data space	A data-sharing architecture in which data remains with the original holder and is accessed through controlled, interoperable and policy-enforced mechanisms.
Identity leakage	The failure mode in which a generative model reproduces identities that are present in its training data.
Mated image set	Two or more face images of the same identity, required to compute genuine (same-person) comparison scores in face-recognition evaluation.
Membership inference	An attack, or the corresponding privacy test, that attempts to determine whether a specific record was part of a model's training data.
Participation factor	The time-bounded percentage with which a metering point participates in an energy community, used to weight measured consumption and generation.
Pseudonymisation	Replacement of direct identifiers with artificial ones such that data can still be attributed to a person using additional information; pseudonymised data remains personal data under GDPR Article 4(5).

Draft

© 2024-2027 CERTAIN

Term, expression	Definition
Re-synthesis (source-anchored)	Full regeneration of a dataset that preserves the structural and statistical properties of a real source ecosystem while creating no correspondence to real persons.
Synthetic data	Artificially generated data that reproduces the statistical and structural properties of real data without containing records of real individuals.

*Table 1-1: **Glossary***

The list of acronyms used in this document and their meaning is explained in Table 1-2 below

Acronym	**Meaning**
AI	Artificial Intelligence
API	Application Programming Interface
AU	(Facial) Action Unit
BFI-2-S	Big Five Inventory-2, Short form
BGBI	Bundesgesetzblatt (Austrian Federal Law Gazette)
BPD	Borderline Personality Disorder
BPFSC-11	Borderline Personality Features Scale for Children, 11-item version
BRS	Brief Resilience Scale
CC	Community Coverage (OBIS measurement family)
CC BY	Creative Commons Attribution licence
CDD	Cooling Degree Days
CERTAIN	Certification for Ethical and Regulatory Transparency in Artificial Intelligence
CNN	Convolutional Neural Network
CP	Community Potential (OBIS measurement family)
CSV	Comma-Separated Values
CTGAN	Conditional Tabular Generative Adversarial Network
CV	Curriculum Vitae

Draft

© 2024-2027 CERTAIN

Acronym	Meaning
CW	Continuous Wave (radar)
CYRM-12	Child and Youth Resilience Measure, 12-item version
DDL	Data Definition Language
DDPM	Denoising Diffusion Probabilistic Model
DER	Distributed Energy Resource
DET	Detection Error Trade-off (curve)
DiCE	Diverse Counterfactual Explanations (framework)
DM	Diffusion Model
DOI	Digital Object Identifier
DPIA	Data Protection Impact Assessment
DVC	Data Version Control
EAG	Erneuerbaren-Ausbau-Gesetz (Austrian Renewable Expansion Act)
EEG	Erneuerbare-Energie-Gemeinschaft (renewable energy community; in this document, not electroencephalogram)
EPC	Energy Performance Certificate
EPEX	European Power Exchange
EU	European Union
EV	Electric Vehicle
FAIR	Findable, Accessible, Interoperable, Reusable
FAR	False Acceptance Rate (biometrics); also used as Financial Asset Recommendation in the finance domain -- see finding C-12
FFT	Fast Fourier Transform
FID	Frechet Inception Distance
FiLM	Feature-wise Linear Modulation
FMCW	Frequency-Modulated Continuous Wave (radar)
FMR	False Match Rate
FR	Face Recognition
FRR	False Rejection Rate

Draft

Acronym	Meaning
GA	Grant Agreement
GAD-7	Generalized Anxiety Disorder-7 questionnaire
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GMM	Gaussian Mixture Model
GP	Genetic Programming
GRU	Gated Recurrent Unit
HDD	Heating Degree Days
HR	Human Resources
ICU	Intensive Care Unit
INCA	Integrated Nowcasting through Comprehensive Analysis (GeoSphere Austria)
IQR	Interquartile Range
ITAS	[technology-affinity scale -- confirm exact expansion with UM]
JSON	JavaScript Object Notation
JSON-LD	JSON for Linked Data
KID	Kernel Inception Distance
KPI	Key Performance Indicator
LFW	Labeled Faces in the Wild (benchmark)
LLM	Large Language Model
LoPF-Q 12-18	Levels of Personality Functioning Questionnaire for ages 12-18
LPIPS	Learned Perceptual Image Patch Similarity
LSTM	Long Short-Term Memory
MAD	Median Absolute Deviation
ML	Machine Learning
MMD	Maximum Mean Discrepancy
NBG	National Bank of Greece

Draft

Acronym	Meaning
nDCG	normalised Discounted Cumulative Gain
OBIS	Object Identification System (metering code standard)
ODbL	Open Database License
OECD	Organisation for Economic Co-operation and Development
OOD	Out-of-Distribution
OPSD	Open Power System Data
OSF	Open Science Framework
PCA	Principal Component Analysis
PHQ-9	Patient Health Questionnaire-9
PII	Personally Identifiable Information
PK / FK	Primary Key / Foreign Key
PSO	Particle Swarm Optimisation
PV	Photovoltaic
RADAR	Radio Detection and Ranging
RCGAN	Recurrent Conditional Generative Adversarial Network
RF	Radiofrequency
RFC	Request for Comments (IETF standard)
RFW	Racial Faces in-the-Wild (benchmark)
RMSE	Root Mean Square Error
ROI	Return on Investment
RoPA	Records of Processing Activities
SDE	Stochastic Differential Equation
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
SQL	Structured Query Language
t-SNE	t-distributed Stochastic Neighbour Embedding

Draft

© 2024-2027 CERTAIN

Acronym	Meaning
TMR	True Match Rate
TRL	Technology Readiness Level
TVAE	Tabular Variational Autoencoder
UK-DALE	UK Domestic Appliance-Level Electricity (dataset)
UMAP	Uniform Manifold Approximation and Projection
UTC	Coordinated Universal Time
UUID	Universally Unique Identifier
VAE	Variational Autoencoder
VOC	Volatile Organic Compound
WMC	Weighted Measured Consumption (OBIS measurement family)
WMG	Weighted Measured Generation (OBIS measurement family)
WP	Work Package

*Table 1-2: **Acronyms***

Draft

© 2024-2027 CERTAIN

1 INTRODUCTION

Synthetic data serves as a practical means of testing and validating AI systems, particularly in cases where real-world data cannot be used due to privacy, legal, or operational restrictions. In this project it is employed to create realistic yet controlled datasets which retain the main features of real data in various domains (Goodfellow et al., 2014 [1] ; Abay et al., 2018 [2]), enabling the partners to reproduce a wide variety of scenarios without having to reveal any sensitive information.

Synthetic data is also well suited to meeting the particular requirements of compliance testing since it allows for a systematic and repeatable assessment of an AI system under controlled conditions, thus helping to identify risks, biases, and performance limitations (Voas & Kuhn, 2021 [3]). The ability to repeat the tests is important when verifying compliance with the European rules on AI and data protection (European Commission, 2021a. European Commission, 2021b [4]), as the same test usually has to be run on many occasions and the results documented.

The document concerned with Task 5.1 outlines the work carried out. Its aims are to produce synthetic test datasets for the various application areas of the project, to record the methods and techniques employed in producing them, and to assist the AI compliance assessment by supplying data that is realistic, diverse and controlled for use in testing. The test data obtained is scalable and comparable across the project because the partners' domain-specific approaches have been combined into a single methodology.

Only if a system reacts to a test dataset in the same way that it would to actual data is it worthwhile to use such a dataset. In all different areas there is a continual need for synthetic data to be not only safe to share but also faithful to the real environment so that the results can be transferred back to it. Each section of the document explains how its data is checked against this standard, and in Section 4, fidelity is made one of the three criteria—along with utility and privacy—by which each partner's method is evaluated.

The work is organised around six application domains, each tied to a project pilot and to a partner that owns the data and the generation method:

- **Finance** (UL and NBG, Pilot 6): the financial asset recommendation has been developed using the FAR-Trans dataset, this being done by incorporating counterfactual scenarios which enrich the market conditions not covered by the historical data.
- **Health** (UM and NVISION): It includes two areas of research: multimodal digital-phenotyping features designed for screening youth mental health (based on the Syn.Mind and SMILE cohorts) and radar-based breathing signals coupled with controlled apnea and motion artefacts.
- **Biometrics** (IPS and Idiap): synthetic face images created for the purpose of evaluating facial-recognition systems, which are designed so that the identities of the real training subjects cannot be traced back.
- **Energy** (SPU and D4P): There are two complementary methods: a re-synthesis of an actual Austrian energy community (EMPOWER) and a generative model which has been trained on public data and can therefore be applied to other countries and conditions.
- **Text** (NVCR): Forums intended for public deliberation which are created by means of simulated agents in order to test discourse-analysis and content-moderation tools.
- **Human resources** (INCO): with regard to HR (Pilot 4), the exception consists in the fact that the limitation in this case is legal access to actual CVs and not a lack of data; therefore, the pilot uses federated, compliance-aware processing rather than creating synthetic datasets.

Draft

© 2024-2027 CERTAIN

1.1 BACKGROUND

As AI systems are used in more critical sectors, reliable testing matters more because a failure can lead to financial loss, safety risks, or broader harm (OECD, 2021 [5]). Whether a test tells you anything depends on the data behind it. Without data that reflect what the system will actually meet, there is no way to judge whether it performs, behaves fairly, or meets the rules (European Commission, 2019 [6]).

Real data is the obvious choice and often the one that's unavailable. GDPR restricts it, owners are reluctant to share it, and in several domains it does not exist at the coverage a good test needs. Synthetic data preserves the statistical properties that make a dataset useful for testing while removing the personal information that would make it impossible to share (El Emam et al., 2020 [7]). That is what makes it suited to controlled, compliant testing - provided it stays close enough to the real data to be a valid stand-in, which is the condition each domain returns to.

1.2 STRUCTURE OF THE DELIVERABLE

The regulatory and legal framework concerning synthetic data is set out in Section 2, the motivations and particular requirements of each domain are given in Section 3, Section 4 forms the main part of the document by describing the method of generation used by each partner and the way in which it was validated, and the conclusions, the references and the supporting material complete the document.

Draft

© 2024-2027 CERTAIN

2 REGULATORY AND LEGAL FRAMEWORK FOR SYNTHETIC DATA

The regulatory and legal framework for synthetic data within the CERTAIN project is designed to translate European legal requirements for artificial intelligence into **operational and testable compliance criteria**, applicable to AI system evaluation in controlled environments. Rather than restating legal obligations in abstract terms, the framework bridges key regulatory instruments (primarily the EU AI Act, the General Data Protection Regulation (GDPR), and fundamental rights law) with the technical methodologies used to test AI systems, including the use of synthetic or simulated data.

Within this approach, synthetic data is understood not merely as a technical tool, but as part of a broader compliance architecture. Its role is to enable testing under conditions that reflect real-world data structures and risks, while reducing reliance on personal data. At the same time, the framework recognises that synthetic data – particularly when derived from real datasets – may still raise legal concerns, including risks of re-identification or unintended replication of sensitive patterns. As such, compliance considerations apply not only to the AI system being tested, but also to the **design and generation of test data itself**.

2.1 KEY TERMS

Personal data means any information relating to an identified or identifiable natural person. Pseudonymisation replaces identifying details with substitute values while a means of re-linking them to the individual continues to exist; pseudonymised data remains personal data. Anonymisation removes identifiability to the point that no individual can reasonably be re-identified by any party, taking into account the means reasonably likely to be used; anonymised data falls outside data protection law. Special category data means personal data revealing racial or ethnic origin, health, sexual orientation, and similar sensitive attributes, which is subject to additional restrictions. Applying a mitigation technique (generalization, noise addition, aggregation) does not by itself achieve anonymisation and whether it does depends on whether residual identifiability risk has actually been reduced to a negligible level.

2.2 LEGAL STATUS OF SYNTHETIC DATA

Synthetic data does not automatically fall outside data protection law merely because it is artificially generated or lacks direct identifiers. Its legal status depends on how it was produced, whether personal data was used as source or training data, and whether individuals remain reasonably identifiable through the generated records, combinations of attributes or linkage with external information. Where personal data is involved, data protection requirements may continue to apply to the generation process even if the final output qualifies as non-personal data. Particular caution is required for rare characteristics, small populations and sensitive or special-category information.

Organisations remain accountable for ensuring that synthetic data is generated and used lawfully. Their assessment should cover the entire data lifecycle, including the lawful basis and purpose of processing, data minimisation, transparency, security, retention and residual re-identification risks. They should also document the source data, methods, assumptions, safeguards and validation results, maintaining sufficient traceability to explain how the dataset was generated and used. Where source data was originally collected for research purposes, the derogations available under Article 89 GDPR (which permit deviation from purpose and storage limitation, subject to appropriate safeguards) may be relevant and should be considered explicitly as part of that assessment, rather than treated as implicit.

Draft

© 2024-2027 CERTAIN

2.3 INTERACTION WITH THE AI ACT

Where synthetic data is used to train, validate or test a high-risk AI system, obligations under the EU AI Act apply in parallel with data protection law. Data governance requirements (Article 10) require training, validation and testing datasets to be relevant, sufficiently representative, and examined for biases likely to affect health, safety or fundamental rights. Article 10(5) permits providers to process special category data, exceptionally and to the extent strictly necessary, for the purpose of detecting and correcting bias in high-risk systems, subject to safeguards including data minimisation, security controls, and deletion once the bias has been addressed.

2.4 RESIDUAL RISKS

Synthetic data may reduce dependence on real-world information, but it does not eliminate privacy, fairness or reliability risks. Three forms of residual risk are particularly important.

Re-identification risk arises where a synthetic record can be linked to a real person. This may result from a rare characteristic, an unusual combination of demographic or contextual variables, or the availability of external datasets that permit linkage. Individually innocuous variables may create substantial identifiability when combined. This is especially relevant to small populations, rare conditions, specialised services or distinctive personal histories.

Memorisation risk arises where a generative model reproduces information from its source or training data. A purportedly synthetic record may replicate a real person's unusual employment history, medical profile or other distinctive characteristics, even though such reproduction was not intended. Similarity testing, assessment of overfitting and systematic review of generated outputs are therefore important parts of validation.

Inference risk arises where synthetic outputs reveal information about the source population or reproduce sensitive associations embedded in the underlying data. For example, a dataset may associate leadership, employment success, health status or criminality with particular demographic groups. Although no individual is identified, the output may perpetuate historical inequalities or disclose patterns concerning a small or vulnerable population.

Residual risk means the risk remaining after reasonable legal, organisational and technical safeguards have been applied; it does not necessarily indicate non-compliance. The objective is to reduce risk to a proportionate and acceptable level while preserving the legitimate scientific or operational value of the dataset. Automatic exclusion of rare or underrepresented populations may itself create ethical and practical harms by making AI systems less reliable for those groups and limiting the ability to test fairness and robustness.

Data subject rights after generation: Where a data subject exercises a right of erasure, rectification or objection over source data, organisations should have a defined process for assessing whether synthetic datasets already generated from that data are affected, and whether regeneration, deletion or other remedial action is required. This should be addressed at the design stage rather than left unresolved until a request is received.

2.5 CORE ETHICAL PRINCIPLES

Legal compliance alone does not guarantee that synthetic data is ethically sound. Artificial datasets may still create or reinforce discrimination, exclude relevant populations, misrepresent human experience or produce

Draft

© 2024-2027 CERTAIN

misleading evidence. Ethical principles should therefore operate as practical criteria throughout dataset design, generation, validation and use.

Fairness and non-discrimination: Synthetic datasets should be examined for direct and indirect bias, including bias introduced through proxy variables such as location, language, education or socioeconomic indicators. Removing a protected characteristic does not necessarily remove its influence. Dataset design should not reproduce historical inequalities or encode stereotypes – for example, by systematically associating leadership with one gender or particular forms of risk with certain ethnic or social groups.

Representation and inclusion. Synthetic data should reflect relevant real-world diversity, including populations and experiences that are frequently absent from conventional datasets. At the same time, inclusion must be balanced with appropriate protection where highly distinctive profiles increase identifiability.

Realism and validity: Synthetic profiles and scenarios should be sufficiently plausible for the intended testing purpose. Implausible combinations of characteristics, artificial correlations or unrealistic distributions may lead to misleading conclusions about an AI system's performance, fairness or safety.

Transparency and explainability: Synthetic data is not inherently neutral or objective. It reflects decisions about source data, variables, distributions, constraints and acceptable deviations from reality. These choices and their limitations should be made visible and documented sufficiently.

Accountability and human oversight: Synthetic data generation should not be treated as a fully autonomous technical process. A designated reviewer with appropriate authority, such as a data protection officer, ethics committee, or equivalent named function, should review and approve generation methodology, validation results and identified residual risks at key stages before the dataset is used.

Proportionality and responsible use: The level of detail, sensitivity and access should be no greater than necessary for the stated purpose.

2.6 APPLICATION TO THE HUMAN RESOURCES PILOT

The human resources pilot is the case where these obligations bite hardest: recruitment systems sit squarely in the AI Act's high-risk category, and the data involved is personal almost by definition.

From the perspective of the AI Act, AI systems used in employment and recruitment contexts are generally considered high-risk, given their potential impact on individuals' access to work and economic opportunities. This classification requires that systems be assessed against criteria such as robustness, accuracy, transparency, and human oversight. Synthetic or controlled datasets can support this assessment by enabling structured testing across a range of scenarios, including edge cases and stress conditions that may not be easily observable in real-world data. In the context of the HR pilot, AI-supported recruitment, CV filtering, candidate recommendation, skills extraction and job matching may fall within the AI Act high-risk category for employment, workers management and access to self-employment. This is particularly relevant where AI systems are used to analyse or filter applications, evaluate candidates, rank profiles or support recruitment-related decision-making. Accordingly, compliance testing should generate evidence relevant to high-risk AI requirements, including risk management, data governance, transparency, human oversight, logging, accuracy, robustness and cybersecurity.

In parallel, the GDPR provides the core framework governing the processing of personal data, which is particularly relevant in the HR domain. While synthetic data can reduce direct exposure to personal data, it does not automatically fall outside the scope of data protection law. The generation process must still be evaluated in light of GDPR principles such as data minimisation, purpose limitation, and security, especially

Draft

© 2024-2027 CERTAIN

where synthetic data is informed by real-world data structures. This ensures that synthetic data functions as a risk-reduction mechanism, rather than introducing new forms of legal or ethical risk. Given the use of AI-supported recruitment, profiling, skills extraction and candidate matching, the HR pilot should include an assessment of whether a Data Protection Impact Assessment (DPIA) is required. Where processing is likely to result in high risk to individuals' rights and freedoms, the DPIA should document the processing purposes, categories of data, lawful basis, necessity and proportionality, risks to data subjects, mitigation measures and residual risks. Where applicable, the assessment should also consider potential impacts on fundamental rights, including non-discrimination, access to employment, human dignity and the right to effective human oversight.

The framework is further grounded in fundamental rights principles, particularly non-discrimination, human dignity, and equal access to employment. These are especially relevant in the HR context, where AI systems may influence hiring decisions or labour market opportunities. Testing methodologies must therefore be capable of identifying both direct and indirect discrimination, including patterns arising from proxy variables such as educational background, language skills, or geographic indicators.

A central feature of the CERTAIN methodology is the operationalisation of these legal principles into concrete **testing criteria**. Concepts such as transparency, robustness, accuracy, and fairness are translated into observable properties of AI systems. Transparency is reflected in documentation, traceability, and audit logs; robustness in system stability and resistance to manipulation; accuracy in context-appropriate performance thresholds; and fairness in structured bias-testing protocols and the assessment of disparate impact.

Finally, the framework places strong emphasis on governance and documentation. Testing outputs are structured to support risk management processes and future conformity assessment, including traceability of model iterations, documentation of testing methodologies, and reporting of bias and robustness metrics. This ensures that testing results constitute **legally meaningful compliance evidence**, rather than purely technical validation outputs. The governance framework should also clarify the legal and operational roles of the actors involved in the federated HR data space. For each processing scenario, the relevant parties should identify whether they act as data controller, joint controller, processor, AI system provider, deployer, data holder or data user. This allocation should be reflected in the applicable agreements, access-control rules, audit procedures and accountability mechanisms.

Example: Synthetic data generation in the HR use case

This example is illustrative in Pilot 4 (Human Resources), compliance testing is primarily conducted through federated and controlled processing of real-world data structures. However, the scenario demonstrates how synthetic data could be incorporated, if required, as a complementary tool within a legally grounded and compliance-oriented testing framework.

To illustrate how the above regulatory framework could be operationalised, a hypothetical scenario can be considered in which synthetic data is used to support compliance testing in the HR domain.

In this scenario, a synthetic dataset would be constructed to reflect the structure and key characteristics of CV data processed within the federated data space. The dataset would consist of artificially generated candidate profiles, designed to represent typical combinations of:

- education and qualifications
- skills and competencies
- professional experience and career trajectories
- job roles and sectoral backgrounds

No identifiable personal data would be included, and profiles would be generated based on generalised patterns rather than real individuals.

To enable meaningful compliance testing, the dataset could incorporate **controlled variation in selected attributes**, such as:

© 2024-2027 CERTAIN

- gender indicators
- educational background (e.g. type or prestige of institution)
- language skills or geographic markers

These controlled variations would allow the system to be tested against key legal risks, in particular:

- whether equally qualified candidates receive different rankings or recommendations
- whether proxy variables indirectly influence outcomes
- whether aggregated outputs reinforce or mitigate labour market biases

The design of the dataset would itself be subject to legal and ethical constraints, including:

- ensuring that no data can be linked back to real individuals (re-identification prevention)
- carefully controlling the inclusion and distribution of protected or proxy attributes
- avoiding the replication or amplification of harmful real-world biases
- ensuring that the dataset supports testing without introducing misleading or unrealistic patterns

In this configuration, synthetic data would not be used for model training, but rather as a **controlled testing layer**, enabling structured evaluation of system behaviour against requirements of:

- fairness and non-discrimination
- transparency and explainability
- robustness under varied conditions

If synthetic or simulated HR data is used, the generation process should be accompanied by privacy and quality validation. This should include checks for re-identification risk, memorisation or replication of real profiles, inappropriate inclusion of protected or proxy attributes, bias amplification and unrealistic attribute combinations. The assumptions used to generate the synthetic profiles should be documented so that the dataset can serve as reliable compliance evidence.

Draft

© 2024-2027 CERTAIN

3 MOTIVATION AND REQUIREMENTS

Compliance testing requires data, and the data that would serve a test best is usually out of reach: either its use is not permitted, or it does not exist in a form that can be shared. An AI system tested for fairness, robustness, or data protection has to run on inputs close to what it will meet in operation, and such inputs tend to carry personal information, fall under legal restrictions, or be missing at the scale a test needs. The usual answer is synthetic data, which can preserve the statistical characteristics relevant for testing while removing the connection to real individuals. A field turns to synthetic data for one of three reasons: the real data is too sensitive to share, there is not enough of it to cover the scenarios a test requires, or it exists but may not legally be made available.

The dominant reason varies from one field to another, which is why the sections that come next are not all the same. In cases where privacy is the main limitation, the approach focuses on replicating the statistical distributions and dependencies in the data without reproducing the individuals who are behind it. When coverage is the limiting factor, the emphasis is on creating scenarios that the actual data never included. If the restriction is legal access, then the question ceases to be one of generation altogether. The form of each subsection is determined by the problem it is addressing rather than by following a set template.

The CERTAIN pilots map onto these reasons as follows:

- Finance: the obstacle is coverage. Public benchmarks span too narrow a range of market conditions, so hypothetical scenarios are generated to fill the gaps.
- Health: the obstacle is sensitivity. Patient recordings cannot be shared at the scale testing needs, so the aim is to reproduce their statistical and physiological characteristics without exposing any individual.
- Biometrics: the obstacle is sensitivity here as well. Biometric data identifies a person by its nature, so evaluation datasets are generated rather than collected, with no traceable link to real subjects.
- Energy: the data sits with operators who are legally unable to make it publicly available, so a real ecosystem is re-created in a shareable form and complemented with data from public sources.
- Text: no shareable record of online deliberation exists, so discussion forums are produced from simulated agents.
- Human resources: the data exists, but legal and ethical restrictions prevent access and sharing, so this pilot defines requirements for compliance-aware federated processing instead of dataset generation.

The same line of reasoning is applied in the subsections below even though their form varies: each one explains why real data is unusable or inadequate in its area and then states the requirements which the method of data generation in Section 4 has to satisfy.

3.1 FINANCE

Financial asset recommendation is a subdomain of recommender systems that aims to identify suitable financial securities for investors, with the expectation that they will invest capital to the recommended assets. Effective recommendation systems rely on modelling pricing data to distinguish between high-performing and potentially undervalued assets. However, the absence of public datasets with realistic customer transactions that can be used to train and evaluate approaches under a common benchmark is a significant barrier. This challenge has been partially addressed by prior work from the NBG which released datasets including pricing time series for multiple asset classes (e.g., stocks, bonds, and mutual funds), asset descriptions, and anonymised customer transactions [8] . Despite this, limitations remain in terms of data diversity and coverage of different market scenarios.

Draft

In this context, synthetic data provides an opportunity to enrich financial time series with diverse and controlled scenarios. By introducing new patterns and variations in trends and volatility, it improves model generalisation and robustness to unseen market behaviours. Beyond augmenting historical data, synthetic data enables the generation of alternative, yet plausible, market conditions.

In particular, counterfactual-based synthetic data offers a powerful way to explore how changes in market conditions affect recommendations [9], addressing the question: "*What would need to change in the market to increase the model's profitability score?*" Since historical data may not fully cover all relevant regions of the decision space, counterfactuals help generate borderline scenarios, improving coverage of decision boundaries and supporting more reliable recommendations. Furthermore, this approach enables the identification of sensitive features, such as volatility and returns, that drive model decisions. By analysing how small changes in these features can alter recommendations, it becomes possible to better understand and stabilise model behaviour. Most importantly, it enhances explainability by providing clear insights into how market conditions would need to change to yield different outcomes.

Table 3-1 Synthetic data use cases and the compliance requirements addressed

Use case	Compliance
Generation of alternative market scenarios (e.g., price shocks, volatility spikes) not present in the historical data to test how the recommendations perform under adverse conditions.	Model robustness and reliability.
Counterfactual explanations for recommendations, showing the minimal change in a client's profile/portfolio (e.g., risk tolerance, holding period) that would lead to a different asset suggestion.	Transparency and explainability.
Personalised recommendation modelling without using real identifiable data by evaluating models on synthetic counterfactual clients instead of real identifiable transaction/customer data.	Privacy preservation via pseudonymisation.
Generate synthetic fraudulent or anomalous transaction patterns to augment the training data set, where examples of real fraud are rare or imbalanced.	Robustness of fraud detection models.

Table 3-2 Risks of synthetic data generation and mitigation measures

Risk	Mitigation
Generated price changes may not reflect realistic market behaviour.	Constrain generation using financial domain knowledge and statistical properties.
Counterfactuals may break time dependencies (e.g., unrealistic jumps in time series).	Enforce temporal coherence constraints and sequence-aware generation methods.
Synthetic scenarios may deviate too far from realistic market distributions.	Use distributional similarity metrics and calibration techniques.
Small unrealistic changes may lead to misleading recommendations.	Enforce stability constraints and validate sensitivity.
Model may prioritise artificial patterns over real dynamics.	Balance training with real data.

Draft

© 2024-2027 CERTAIN

Generating high-quality counterfactuals can be expensive.	Use efficient approximation methods and sampling strategies.
The public availability of artifacts provides anyone with the ability to create adversarial or poisoning attacks against similar production systems.	Advise future users to validate the robustness of the system before using it in production.
Malicious external contributions (e.g., pull requests, updated datasets) could reintroduce backdoors or biased counterfactuals.	Require review/approval for all external contributions.

3.2 HEALTH

Over the last few decades, digital healthcare has increasingly moved toward predictive, preventive, personalised, and participatory (P4) approaches [10], a shift also supported by advancements in AI systems that draw on behavioural, physiological, and contextual data. However, the development and assessment of such systems are often constrained by the limited availability of representative data, strict privacy requirements, and the need to demonstrate robustness, fairness, and regulatory compliance. Synthetic data therefore provide an important mechanism for enabling trustworthy AI testing while reducing barriers to data access and sharing. Within the health pilot considered in this project, synthetic data support complementary aspects of mental health assessment and monitoring, ranging from multimodal digital phenotyping to physiological signal analysis. These activities illustrate how synthetic data can facilitate development, validation, and compliance testing in sensitive healthcare settings.

Digital Phenotyping for Mental Health Assessment

Screening for mental distress in youth increasingly requires approaches that move beyond symptom checklists toward the systematic quantification of behavioural and physiological correlates of anxiety and depression captured during standardized assessments. Although established instruments such as structured clinical interviews, the Generalized Anxiety Disorder-7 (GAD-7), and the Patient Health Questionnaire-9 (PHQ-9) provide validated indices of symptom burden, they remain constrained by self-report bias, clinician-dependent variability, and limited scalability in routine and research settings. In this context, digital phenotyping offers a complementary framework in which observable cues, such as, linguistic content and style, speech prosody, facial affect, gaze and head movements, motor behaviour, contactless cardio-respiratory signals, and environmentally derived indicators—are operationalized as measurable multimodal features associated with anxiety- and depression-related symptom dimensions. Integrating these features with questionnaire scores and available clinical labels enables the derivation of candidate digital biomarkers, supports interpretable risk modelling, and provides the basis for privacy-preserving synthetic data generation for downstream methodological development.

At the same time, the collection of high-resolution video and audio recordings of vulnerable adolescents raises substantial concerns regarding identifiability and secondary use, which are addressed by a privacy-by-design strategy in which raw recordings are processed locally, deleted after feature extraction, and never shared externally. Synthetic data are thus central to reproducibility and open science. They enable broader reuse of rich, clinically annotated multimodal signals without exposing identifiable video, audio, or high-dimensional behavioural traces from individual adolescents. In the context of digital phenotyping, generative models can be trained on extracted feature streams (for example gaze dynamics, facial affect trajectories, voice prosody, interaction tempo, and contactless physiological markers) to produce artificial samples that preserve key joint distributions and temporal dependencies relevant for anxiety and depression risk, while breaking the direct link to any specific participant. Such synthetic datasets can then be shared more openly for benchmarking, method development, and external validation of explainable AI models, helping to address current bottlenecks in reproducibility and generalizability that arise from small, siloed, and legally constrained clinical datasets.

Draft

© 2024-2027 CERTAIN

The initial Syn.Mind cohort, captured in CERTAIN project, constitutes the empirical basis for this work as an observational, cross-sectional pre-screening study conducted at the University Medical Centre Maribor in adolescents and young adults aged 12–20 years. The cohort includes participants with anxiety disorders, depressive disorders, borderline personality disorder, and non-clinical controls, and uses a 2 × 4 design with random allocation to one of two standardized interview modalities: a text-based chatbot or a virtual-agent interface delivered on a tablet device. During a single assessment visit, participants complete baseline questionnaires, including GAD-7 and PHQ-9, a structured digital pre-screening interview, and post-interview acceptability measures, while multimodal data are recorded through audio-video capture, contactless physiological monitoring, and continuous indoor air-quality sensing. The resulting dataset combines demographic and clinical information, questionnaire responses, interview-experience outcomes, and derived linguistic, speech, facial, movement-related, physiological, and environmental markers, which are used both for primary digital-phenotyping analyses and for carefully governed secondary uses. In line with European data-protection and open-science requirements, secondary use explicitly excludes sharing of raw video or audio; instead, access for the broader research community is envisioned via privacy-protected feature sets and synthetic multimodal datasets that approximate the joint distributions and temporal dynamics of the original recordings without exposing identifiable visual or vocal information.

To sum-up for the Syn.Mind cohort, synthetic data generation thus functions as a privacy-preserving bridge between a highly sensitive primary dataset and the need for scalable, transparent, and independently verifiable models of mental distress in youth.

Table 3-3 Health, digital phenotyping: use cases and compliance requirements

Use case	Compliance
Generation of synthetic multimodal feature streams (speech, facial, movement, physiology, environment) reflecting anxiety and depression risk patterns across diagnostic groups	Representational fairness across diagnoses, gender, and age bands
Creation of noise-augmented synthetic sequences to simulate missing data, sensor noise, and motion artefacts in audio-video and physiological channels	Model robustness to real-world recording conditions and technical variability
Generation of synthetic feature vectors at interview-window level for external benchmarking of explainable AI models	Reproducibility and comparability of methods without exposing identifiable raw video/audio
Production of strongly de-identified synthetic datasets for sharing under open-science frameworks (for example via Zenodo)	Compliance with data-protection rules and restrictions on secondary use of sensitive youth psychiatric data
Simulation of rare but clinically important configurations (for example high symptom burden with specific cue patterns) to support model calibration	Mitigation of class imbalance and improved performance on under-represented clinical profiles

Table 3-4 Health, digital phenotyping: risks and mitigation measures

Risk	Mitigation
Synthetic multimodal feature distributions (for example speech, facial, movement, physiology)	Use distributional similarity metrics and calibration against real Syn.Mind feature distributions stratified

Draft

© 2024-2027 CERTAIN

may deviate from realistic patterns observed in adolescents with and without anxiety or depression.	by diagnosis, gender, and age band before releasing synthetic datasets.
Generated sequences may not preserve clinically meaningful associations between observable cues and symptom severity (for example GAD-7/PHQ-9 scores).	Constrain generative models using joint statistics of cues and symptom measures; validate on held-out real data using traditional ML models (for example logistic regression, random forests) and PSO-based feature selection to check that risk-related features and effect sizes remain stable.
Severe class imbalance (for example fewer high-severity or BPD cases) may lead to biased models when synthetic data are used together with real data.	Combine synthetic generation with classical imbalance techniques (for example SMOTE-like oversampling and class-balanced loss functions), and systematically compare performance and feature importance between models trained with and without synthetic augmentation.
Noise augmentation and artifact simulation may produce unrealistic patterns or overly severe distortions in audio-video and physiological channels.	Bound noise-injection parameters using empirically observed artefact characteristics in Syn.Mind recordings, and conduct expert review plus quantitative plausibility checks on synthetic samples.
Synthetic data may inadvertently encode spurious site- or device-specific artefacts (for example room setup, sensor placement) that limit generalizability.	Apply normalization and domain-adaptation steps before model training; evaluate logistic-regression and tree-based models across technical subgroups, and use PSO to down-weight or exclude features that primarily capture site/device artefacts.
Residual re-identification risk from high-dimensional behavioural signatures, even when raw video and audio are not shared.	Restrict sharing to feature-level and aggregate synthetic datasets, apply privacy checks (for example nearest-neighbour distance to real samples and disclosure-risk metrics), and keep raw recordings local and deleted after feature extraction.

Contactless Physiological Signal Monitoring

Radiofrequency-based (RADAR) contactless physiological monitoring has been gaining momentum in Digital Health, and more specifically in sensor-enabled patient monitoring. Affordable hardware radar kits have made the technology accessible to research groups and industry alike, enabling a monitoring approach that offers meaningful advantages over competing solutions: it requires no physical contact, does not interrupt daily activities, operates independently of lighting conditions, can record through clothing, and can be unobtrusively integrated into everyday environments.

Despite these promising properties, the broader adoption of radar-based monitoring in clinical and healthcare settings remains constrained by a fundamental data challenge. Unlike the well-established tradition of medical-grade physiological datasets, or the rapidly growing body of data driven by the wearable health market, radar physiology datasets have only begun to emerge in the last five years. Data scarcity remains a critical obstacle: available datasets are limited in size and diversity, and the signals they contain are frequently affected by motion artifacts, introducing noise that compromises model reliability. Furthermore, existing datasets exhibit demographic gaps. For instance, apnea-related data has been collected predominantly from adult subjects, despite the fact that breathing patterns vary substantially across age groups. This introduces a risk of bias in AI models trained on such data, limiting their generalizability and clinical trustworthiness.

These conditions create both a barrier and an imperative. Poor data quality and demographically unrepresentative datasets can undermine external validation, produce dataset shift and subgroup performance

Draft

disparities, and limit the robustness and generalizability of AI systems across intended patient populations, making it difficult to demonstrate the robustness, fairness, and regulatory compliance required for adoption in healthcare [11]. Synthetic data generation may help address gaps in data availability and representation, provided that the resulting data are rigorously evaluated.

In this context, the proposed methodology targets two specific requirements. First, it aims to produce high-fidelity synthetic breathing signal datasets that can serve as controlled testbeds for AI model evaluation. By varying clinically and non-clinically relevant signal characteristics, these datasets may be used to complement real-world data to support model development, expose subgroup-specific weaknesses, and improve robustness. Second, the methodology provides a set of complementary metrics for assessing the generated data

More concretely, this entails synthesising breathing signals across age groups that are underrepresented in current data and producing noise-augmented signals that replicate the motion artefacts inherent to radar acquisition, enabling the systematic stress-testing of AI models under realistic operating conditions.

This motivation is particularly relevant to predictive, preventive, personalized, and participatory (P4) medicine, whose implementation depends on integrating heterogeneous data and developing models that remain reliable across diverse individuals and clinical contexts. However, the reality for specific health applications is that real-world data often suffers from scarcity, sparsity, or unbalance problems which complicates model development and limits rigorous validation [12]. The present work therefore focuses on developing and evaluating a controlled synthetic breathing-signal framework, rather than claiming to reproduce the full diversity of real-world populations. Its purpose is to provide an additional tool for model development and methodological evaluation, to be used alongside appropriately designed validation on real-world data.

Table 3-5 Health, radar monitoring: use cases and compliance requirements

Use case	Compliance
Generation of signals with apnea patterns across age groups	Representational fairness
Noise-augmented synthetic signals to produce motion-like artifacts.	Model robustness

Table 3-6 Health, radar monitoring: risks and mitigation measures

Risk	Mitigation
Synthetic age-group distributions may deviate too far from realistic physiological variation.	Use distributional similarity metrics and calibration against available reference data.
Generated breathing signals may not reflect realistic physiological patterns.	Constrain generation using domain knowledge of respiratory physiology and statistical properties of real radar signals.
Noise augmentation may produce artifact patterns that are unrealistic or too severe.	Bound noise injection parameters based on empirically observed motion artifact characteristics in radar signals.
Synthetic apnea events may not capture clinically meaningful pattern variability.	Extract physiological features from generated signals and compare them against clinically established parameters from the literature

Draft

© 2024-2027 CERTAIN

3.3 BIOMETRICS

In the Biometrics pilot, we will test an identity-verification system for secure access using Facial Recognition (FR) technology. (For the pilot, the FR system will be provided by IPS, and will be evaluated by IDIAP.) In general, FR algorithms and products are, of course, thoroughly evaluated, both in-house and via independent benchmarks [13] [14], before being delivered to clients. Nevertheless, it is still necessary to evaluate systems *in-situ*, both as a sanity check and to avoid problems related to the client's specific infrastructure.

FR systems typically capture a live face-image (or video) sample called (a *probe* sample), and compare it to a *reference* sample. The reference sample may have been recorded previously, or may also be captured at the same time as the probe sample. For example, in a border control scenario, a traveller may first present his or her passport (from which the reference data is scanned), and then provide a live probe sample via the camera in the *automatic border control* (ABC) gate. The FR system will then call a service -- via an API -- to compare the reference and probe samples. Synthetic datasets can be very useful for testing such FR systems.

Whereas *online testing* of the full system – including both the acquisition system and API – is valuable, it is difficult to perform online tests of biometrics systems in sufficiently large volume, for a robust evaluation. Also, online testing may only be performed at the system level, not at the algorithm level. Therefore, we often resort to *offline testing*, where we bypass the acquisition stage and use an appropriate pre-recorded dataset, to perform a robust evaluation of the core biometrics functionality using the provided API. Using sufficiently large datasets in offline testing scenarios enables us to quantify the FR performance much more accurately and precisely. This also enables us to test the throughput, by simulating busy scenarios, which would be difficult to do in live (online testing) scenarios. The evaluation dataset should meet the following requirements.

Requirements:

- **Mated sets.** It should contain mated image sets, that is at least 2 images per identity, to be able to perform evaluation using standard FR metrics.
- **At least 7747 identities.** The test dataset should contain samples of at least 7747 identities. This allows calculation of at least 30,004,131 impostor scores and reliable estimates of performance at operating points down to one error in a million (FAR=10^{-6}).
- **Diverse.** It should represent a diverse cross-section of the public in terms of gender, age and ethnicity.
- **References and Probes.** Each mated image set should contain both reference and probe images. Reference images should represent a controlled capture scenario with neutral pose, expression, illumination and background. Probe images, on the other hand, should be representative of live capture scenarios, that is, they should contain a variety of poses, expressions, illumination conditions and backgrounds
- **GDPR-compliance.** The dataset should be free of usage restrictions, that is, it should be appropriate to transmit the dataset to a client site and to process it using the client's servers. Ideally, it should also be possible to store the dataset indefinitely.

To form such an evaluation dataset, there are essentially three options: 1) Use a publicly available dataset; 2) Invest in acquisition of an appropriate dataset ensuring purchase of full rights to the images; 3) Generate an appropriate synthetic dataset. Table 3-7 provides an assessment of these three options in terms of our requirements.

Table 3-7 An assessment of potential solutions for evaluation dataset formation.

Dataset type	Requirements				
	Mated sets	**At least 7747 IDs**	**Diverse**	**References & Probes**	**GDPR-compliant**

Draft

Public (in the wild)	Yes	Yes	Yes	Typically, no reference images are available	Yes, although individuals may request images to be removed
Public (controlled)	Yes	Maybe	No	Yes	No
Acquired	Yes	Expensive	No	Yes	Yes, although individuals may request images to be removed
Synthetic	Yes	Yes	Yes	Yes	Yes

Whereas public datasets are a reasonable alternative, datasets containing images of real subjects can never be 100% GDPR-compliant. Even if care is taken to procure extensive rights to the images, subjects are still at liberty to revoke these rights, and so datasets may not perdure. For real datasets, we also see a compromise between the presence of controlled reference images and diversity: enforcing controlled acquisition necessarily limits the data sample to subsets of the population, and conversely, where diverse sets of images are harvested on-line, controls are generally not implemented. Synthetic datasets reap the benefits of generation methods, learning from both controlled and in-the-wild training datasets: identity-diversity is learned from public, in-the-wild datasets while the style of reference images can be learned from controlled datasets.

3.4 ENERGY

Our contribution presents a single, unified methodology for the synthetic generation of high-fidelity energy data. It addresses the need for large-scale, non-sensitive, and operationally realistic energy datasets, which are essential for driving digital innovation in European energy markets and for the algorithmic compliance testing that this deliverable targets. The shared motivation is that the energy data most useful for training and evaluating AI systems is also the most difficult to obtain. Real smart-meter data is highly granular and attributable to identified natural persons, so it cannot in practice be opened for research even when operators wish to do so, while large-scale and cross-country data that satisfies physical and regulatory constraints remains scarce. Our methodology overcomes these limitations by producing a physically plausible and privacy-safe simulation of complex energy environments.

To cover the full spectrum of energy dynamics, the framework operates through two deliberately distinct and complementary tracks.

1. The first is a source-anchored track that re-synthesises a real, regulated Austrian energy-community ecosystem.

2. The second is a generative track that learns from a geographically diverse pool of public datasets.

The two tracks are complementary rather than redundant. The source-anchored track provides a realistic, in-distribution Austrian benchmark, whereas the generative track extends coverage to households, grids, and conditions that lie beyond any single source.

The source-anchored track is grounded in the EMPOWER citizen-energy-community ecosystem, which is a coherent set of real Austrian energy communities (Erneuerbare-Energie-Gemeinschaften, EEG) operating under a single regulatory frame defined by the Renewable Expansion Act (Erneuerbaren-Ausbau-Gesetz, EAG, BGBl 150/2021). As the academic partner within this ecosystem, we are able to ground a full re-synthesis in a real ecosystem and to produce an openly publishable, GDPR-compliant dataset that preserves the specific regulatory and operational mechanics of Austrian energy communities. This track is required to reproduce the structure of a regulated energy community rather than generic load profiles. It must represent the community entities and their time-bounded relations, namely energy communities, participants, and metering points together with their membership and ownership over time, and it must model participation factors as first-class, time-varying data. Measurements must follow the OBIS-aligned families that govern community accounting, which are weighted measured consumption, weighted measured generation, weighted surplus generation,

Draft

community coverage, and community potential. These values are expressed in kilowatt-hours, are non-negative, and are recorded at a 15-minute, timezone-aware resolution. The track covers photovoltaic and small-hydro generation, while wind is excluded because the source contains too few wind metering points to meet the dataset's k-anonymity threshold. The data must also remain spatially coherent with the Austrian source region, so that external reference data, namely weather and day-ahead market prices, can be linked through geographic coordinates.

The generative track applies advanced machine learning techniques, and in particular conditional diffusion architectures, to generate highly realistic energy time series that generalise behavioural patterns across different countries, climates, and household types. It is designed to produce rich synthetic data that includes timestamps, household information such as building energy class, country, weather conditions, and household-level energy consumption. In addition, it generates data on energy production and consumption at the aggregated grid level across different countries. A key requirement for this track is that the generated data satisfies important physical and mathematical constraints, so that the results remain accurate, valid, and grounded in real-world factors such as regional weather patterns and total energy demand. The track must capture complex temporal dependencies, seasonal trends, and human behaviour patterns such as daily routines and seasonal heating or cooling needs, and it must correctly learn the causal relationships between environmental factors, for example weather conditions, and energy consumption.

Taken together, the two tracks satisfy a common set of requirements in terms of both content and fidelity. On the content side, the unified dataset captures multi-scale energy information that spans the household level, the aggregated grid level, and the regulated energy-community level, together with the weather and market-price context that drives it. On the fidelity side, the synthesis must preserve the structural and statistical properties of real systems, including temporal autocorrelation, cross-series correlation, and the coupling between energy series and local weather, while containing no personal data. By faithfully replicating both the stochastic behaviour of individual consumers and the rigid constraints of regulated energy markets, the contribution provides a robust in-distribution and out-of-distribution substrate for stress-testing and certifying downstream AI forecasting systems prior to deployment.

3.5 TEXT

The development of the Public Deliberation Synthetic Datasets Generation Component responds to a clear and pressing need within the research community: the ability to validate, test, and refine AI systems that engage with online discourse, without the legal and ethical complications that arise when working with real human data. Collecting genuine conversation data from civic platforms or public forums introduces a range of risks, including potential violations of data protection regulations such as the GDPR, difficulties in obtaining informed consent at scale, and significant constraints on how the data may subsequently be shared or reused. The Public Deliberation Synthetic Datasets Generation Component addresses all of these concerns by generating entirely artificial datasets that simulate realistic online deliberation dynamics whilst maintaining complete reproducibility and eliminating privacy risks at the outset.

Three principal requirements underpin the system. First, there is a need for safe, reproducible test datasets that allow developers and researchers to evaluate AI systems for bias, robustness, and performance without risking privacy violations or non-compliance with EU data protection law. Second, the system must support diverse and controlled scenarios, enabling researchers to examine how AI tools respond across a range of participant demographics and discussion contexts. Third, the datasets must be freely shareable and reusable across the research community without the restrictions that real-world data would inevitably impose.

The motivation for each generation method follows directly from these requirements. The agent-based simulation approach is necessary to produce behaviourally plausible conversations, where posting patterns, reply structures, and participation rates reflect authentic online deliberation dynamics rather than artificially uniform behaviour. The AI-driven language generation method is required to produce natural-sounding, contextually coherent text that would be indistinguishable from genuine civic discourse, whilst remaining entirely synthetic. Finally, the procedural generation of demographic metadata is motivated by the need to

Draft

represent realistic population diversity — in terms of age, nationality, and occupational background — without any correspondence to real individuals.

Taken together, these methods fulfil the core requirement of providing a privacy-by-design testing infrastructure that supports the EU AI Act's emphasis on transparent, auditable, and responsible data practices.

3.6 HUMAN RESOURCES

The Human Resources (HR) domain within CERTAIN is characterised by the use of sensitive personal data, including CVs, professional profiles, and employment histories. Unlike other domains, the primary constraint is not the lack of data, but rather the **legal and ethical limitations on accessing, sharing/publishing, and processing such data**.

The motivation for the HR pilot is therefore not the generation of synthetic datasets, but the **development and validation of compliance-aware mechanisms** for:

- Federated data access and search across distributed HR systems
- AI-supported recruitment decision making and workforce analytics
- EU regulatory requirements, GDPR-compliant data sharing and consent management
- Fairness and non-discrimination in AI-driven decision-making

The co-creation process further emphasises the need to ensure that the dataspace and integrated AI tools are aligned with real stakeholder needs, regulatory requirements, and operational practices.

In this context, the HR domain serves as a **high-risk, high-relevance use case** for testing how AI systems can operate under strict compliance constraints without relying on centralised or synthetic datasets.

Requirements

Given this context, the HR domain defines a set of functional, legal, and technical requirements for compliance testing:

Legal and regulatory requirements.

- Full compliance with GDPR principles, including:
  - anonymisation and data minimisation
  - purpose limitation
  - documented lawful basis for each processing purpose
  - consent management
  - consent management, where consent is the appropriate legal basis
  - mechanisms for withdrawal of consent and respect of data subject rights, where applicable
- Prevention of unauthorised access to personal data in federated environments
- Ensuring auditability and traceability of AI-supported decisions
- Ensuring that AI-generated rankings, recommendations or matching outputs remain subject to appropriate human review
- Avoiding solely automated recruitment decisions with legal or similarly significant effects unless the applicable legal conditions and safeguards are met
- Providing mechanisms for explanation, contestability and correction where AI-supported outputs affect candidates or freelance experts
- Alignment with emerging AI regulatory requirements (e.g. fairness, transparency, human oversight)

Draft

© 2024-2027 CERTAIN

- Assessment of DPIA requirements for AI-supported recruitment, profiling and candidate matching
- Assessment of potential fundamental-rights impacts, including risks to non-discrimination, equal access to employment and human dignity
- Documentation of mitigation measures and residual risks

Technical requirements

- Support for federated data processing, where data remains with the original holder
- Interoperability across heterogeneous HR systems and CV platforms
- Secure identity and access management mechanisms
- Integration of RegOps workflows for automated compliance enforcement
- Capability to test AI tools without exposing raw personal data

Operational Requirements

- Usability for multiple stakeholder groups (project managers/recruiters, job seekers/freelance experts, policy makers, AI system providers and Dataspace developers)
- Support for real-world recruitment workflows (publishing, searching, matching)
- Mechanisms for stakeholder feedback and iterative refinement through co-creation
- Compatibility with existing HR platforms and data ecosystems

Approach (in lieu of synthetic data generation)

As synthetic data generation is not performed in this domain, the HR pilot relies on a **combination of alternative approaches**:

- Use of **anonymised or pseudonymised data** where necessary
- Simulation of federated data interactions across distributed nodes
- Testing of AI systems within controlled environments (mock operationalisation)
- Use of scenario-based and workflow-based evaluation rather than data augmentation

This approach ensures that compliance testing remains **realistic, legally valid, and operationally meaningful**, while avoiding the risks associated with generating or using artificial datasets.

Table 3-8 Compliance mapping

Use case	Compliance requirement addressed
Federated CV search	GDPR-compliant data access and data minimisation
AI-driven candidate ranking	Fairness, non-discrimination, and accountability
Skills extraction and profiling	Transparency and explainability
Cross-platform data sharing	Data sovereignty and interoperability
Workforce analytics and policy insights	Auditability and lawful processing
AI-supported CV filtering and candidate recommendation	AI Act high-risk evidence, human oversight, transparency and accountability

Draft

Scenario-based recruitment testing	Fairness, non-discrimination, robustness and bias detection
Audit logs and matching-output traceability	Logging, monitoring, auditability and post-deployment accountability
Synthetic or simulated HR profiles, where used	Re-identification prevention, bias control and synthetic-data validation

Table 3-9 Risks and mitigation measures

Risk	Mitigation
Unauthorised access to personal data	Federated architecture and strict access controls
Bias in AI-driven recruitment decisions	Fairness monitoring and evaluation mechanisms
Lack of interoperability across HR systems	Standardised data formats and APIs
Insufficient transparency of AI outputs	Explainability tools and documentation
Regulatory non-compliance	Integration of RegOps workflows and legal constraints
Overreliance on limited datasets	Scenario-based testing and stakeholder validation
Unclear allocation of legal responsibility between HR actors	Define controller, processor, joint-controller, provider and deployer roles for each processing scenario
Overreliance on consent as a lawful basis	Conduct lawful-basis assessment per processing purpose and document the rationale
Solely automated or insufficiently reviewed recruitment decisions	Maintain human oversight, review procedures, explanation mechanisms and contestability channels
Re-identification or replication risk in simulated/synthetic HR data	Apply privacy validation, re-identification testing and bias-amplification checks
Excessive retention of CV data, logs or AI outputs	Define retention periods and deletion/anonymisation procedures
Inadequate information to candidates or workers	Provide clear notices explaining AI-supported processing, purposes, rights and available safeguards

Draft

4 GENERAL METHODOLOGY FOR SYNTHETIC TEST DATA GENERATION

This section documents how each domain generated its synthetic test data. The subsections share a common structure: the domain and the AI systems under test, the characteristics of the source data, the generation methodology and its constraints, the intended use for compliance testing, and how the outputs are shared under the FAIR (Findable, Accessible, Interoperable, Reusable) principles. The techniques differ because the data differs. Finance applies counterfactual generation to financial time series (Section 4.1), while health combines tabular and time-series generative models for clinical features with conditional generation of radar breathing signals (Section 4.2). Biometrics synthesises face image datasets with GAN- and diffusion-based pipelines (Section 4.3). Energy runs two complementary tracks: a source-anchored re-synthesis of a real Austrian energy-community ecosystem and diffusion-based generation from public datasets (Section 4.4). Text builds deliberation datasets through agent-based simulation with LLM-generated content (Section 4.5). Human resources (Section 4.6) keeps the same structure but has no generation methodology, since compliance testing in that domain relies on federated processing of real data rather than on synthetic datasets.

4.1 FINANCE

4.1.1 Domain Description

Financial Asset Recommendation systems play a key role in the finance domain supporting investment decisions by identifying suitable financial securities based on pricing data and asset characteristics. The domain is characterised by time-dependent financial data, primarily in the form of price time series and historical transaction data.

In such systems, the primary objective from the customer's perspective is to maximise profit. Consequently, evaluation strategies focus not only on predictive accuracy but also on financial performance. Common evaluation metrics include Return On Investment (ROI), which quantifies the potential gains or losses associated with following a recommendation, as well as technical indicators that capture key performance aspects of assets over a given time horizon.

In this context, a profitability-oriented model is developed based on a regressor to predict ROI over a six-month period. This approach enables the assessment of asset recommendations in terms of expected financial return, aligning model outputs with real-world investment objectives.

4.1.2 Source Data Characteristics

A financial asset recommendation dataset (FAR-Trans) collected from a large European financial institution is used for this purpose. This tabular dataset represents a 5-year snapshot of the Greek market, and covers a range of different assets: stocks, bonds and mutual funds for the period between January 2018 and November 2022, inclusive. In addition to asset pricing data for that period, the dataset also includes investment transaction logs (asset buy and sell actions) handled by the institution. The dataset is available at https://doi.org/10.5525/gla.researchdata.1658 .

The investment interactions between banking customers and financial assets represent acquisitions and sales of the securities by individual investors which have been managed by the financial institution. Investment transactions can be used to model the past behaviour of customers and develop personalised recommendation approaches. In the end, we have 806 assets in which 388,049 transactions are available, corresponding to 29,090 customers.

For the profitability-based model, which consider the pricing history algorithm of the assets, we craft a selection of 30 technical indicators (KPIs) based on close price as features: average price, ROI, volatility, moving average convergence divergence, momentum, rate of change, relative strength index, detrended close

Draft

© 2024-2027 CERTAIN

oscillator, ROI/volatility ratio, and maximum and minimum values over a time period prior to prediction. Target (profitability) is also computed, which corresponds to the future return considering 21 trading days. All these features are in the form of *floating-points*.

4.1.3 Synthetic Data Generation Methodology

Counterfactual-based synthetic data generation relies on an already trained predictive model. In our setting, this corresponds to a trained regressor that estimates the profitability of investing in assets. The predictor takes as input price time series, from which relevant KPIs are computed as a preprocessing step. Since financial markets exhibit temporal volatility, this problem requires the use of expanding window experiments.

Given the large number of possible model configurations, we select only the final window from each experimental setting, specifically: 2019-08-01 to 2020-08-28 and 2020-09-14 to 2021-11-23. This results in counterfactual generation at two distinct points in time, corresponding to two synthetic datasets.

We generate counterfactuals using a genetic algorithm from the DiCE (Diverse Counterfactual Explanations) framework [15], with additional domain-specific constraints applied. In particular, only the asset's own historical windows prior to the query date are considered valid reference points. Counterfactuals are accepted only if they satisfy both a minimum and maximum constraint. The minimum threshold is defined as the factual prediction plus a small ε value, ensuring that the counterfactual must strictly improve over the current predicted outcome; this epsilon serves only to avoid numerical ties, so that a counterfactual predicted equal to the factual is not mistakenly accepted as an improvement. The maximum threshold corresponds to the best predicted performance that the asset has achieved in its historical windows (or, if the asset has no prior history, the best performance observed across all assets). This upper bound restricts the search to historically achievable values. Instead of allowing the optimiser to pursue arbitrarily large, unconstrained ROI improvements that the regressor was never trained to support and cannot reliably extrapolate to, the target is capped at a value the asset is known to be capable of reaching based on precedent. Therefore, the generated counterfactual must lie within this feasible range: it should improve upon the current state without exceeding the asset's historical best, keeping the result both meaningful and realistic

In addition, feasibility constraints for genetic algorithm are enforced on the reference data: each counterfactual must be supported by at least one reference window satisfying the desired minimum threshold. This means that a single counterfactual is generated per query, since a well-optimised alternative is more actionable than a diverse set of alternatives. Also, there must be at least 10 unique historical windows in total to produce a single counterfactual. Such a requirement is due to DiCE's genetic algorithm which initialises a population of size 10 for its search. Ensuring that there are at least 10 unique reference windows guarantees that the genetic algorithm can be seeded with a non-degenerate initial population.

To evaluate the quality of the generated counterfactuals, multiple standard quality metrics were considered. Validity was used to verify whether the counterfactual successfully achieved the desired prediction outcome. Lift measured the magnitude of improvement compared to the original factual prediction. Distance-based metrics such as L1 and L2 distances quantified how much the feature values changed between the factual and counterfactual instances, with lower values indicating smaller and more realistic modifications. Mean absolute difference was also used to assess the average feature-level change, while sparsity measured how many features were modified, favouring explanations that required fewer changes and were therefore easier to interpret and act upon. Together, these metrics ensured that the generated counterfactuals were not only valid in achieving better predicted outcomes, but also realistic, interpretable, and practically useful for asset recommendation decisions.

4.1.4 Intended Use for Compliance Testing

There are several distinct ways that generated counterfactuals can be used.

1. **Explainability and Transparency**: The primary goal is to understand and explain why specific assets are recommended to particular customers. This is a direct regulatory requirement under GDPR Article

© 2024-2027 CERTAIN

22 and the EU AI Act Article 13 (transparency obligations for high-risk AI). Counterfactual explanation methods allow to answer the question: "What would need to change in a customer's profile for a different recommendation to be made?"

2. **Fairness**: Verify that the model does not systematically produce different recommendations for different demographic groups (e.g. age, income level, risk profile category). This is required under the EU AI Act's high-risk AI classification applicable to investment recommendation systems.
3. **Profitability vs. Transaction Alignment**: Assess whether the assets recommended by the model are aligned not only with customer transaction patterns (transaction-based evaluation, nDCG) but also genuinely profitable (profitability-based evaluation, ROI).
4. **Robustness**: Evaluate the stability of the model across different market conditions. In particular, consider whether the quality of recommendations is maintained during market downturns as well as during stable or growth periods.
5. **Privacy and Data Minimisation**: Evaluate whether the FAR-Trans model preserves privacy and complies with data minimisation principles under different scenarios. In particular, consider whether privacy guarantees are maintained during training, inference, explanation generation and testing.
   - Model level: The FAR-Trans training data should not enable re-identification of NBG customers, and the model should not memorise sensitive patterns from training data.
   - Data pipeline level: During inference, only the minimum necessary customer data should be processed (data minimisation); decision logs should not store more data than required.
   - Counterfactual explanations: The counterfactuals must not reveal data belonging to other customers, and explanations provided to customers must respect third-party privacy.
   - Testing and evaluation: Trustworthiness testing uses synthetic data — not real customer data — ensuring privacy is preserved throughout the evaluation process.
   - Certification: The certification process (WP6) should document that the AI system implements privacy by design and privacy by default (GDPR Article 25).
6. **Regulatory Compliance Readiness**: Overall, the evaluation should demonstrate that the investment recommendation AI system can meet the conformity assessment requirements of the EU AI Act for high-risk AI systems, supporting the pilot's TRL advancement from 5 to 6.

4.1.5 FAIR Principles for Data Sharing

The generated datasets are publicly available on GitHub (https://github.com/serval-uni-lu/counterfactualFAR), together with the full source code used in the experiments. The repository also includes a requirements file specifying the exact library versions and dependencies used, ensuring reproducibility of the experimental setup. The datasets are provided in CSV format, which supports interoperability and facilitates their use across different tools and machine learning environments. Detailed descriptions of all features and variables are available in the original repository documentation, allowing users to correctly interpret and reuse the data. This setup supports the FAIR principles for data sharing by improving findability through public repository access, accessibility through open availability, interoperability through standardized CSV formatting, and reusability through clear documentation, version control, and reproducible code.

4.2 HEALTH

The health domain considered in this task encompasses heterogeneous data sources that capture complementary aspects of human behaviour and physiology of CERTAIN health pilot described in Section **Errore. L'origine riferimento non è stata trovata.**. The following subsections describe the domain (4.2.1), the source data (4.2.3), the synthetic-data generation methodologies (4.2.4), their intended use for compliance testing (4.2.5), and the FAIR data-sharing principles adopted (4.2.6). They provide a coherent methodological foundation for the generation and evaluation of synthetic health-related data. To improve readability, the following sections are organised around the two complementary aspects of the health pilot with subsections for each topic.

© 2024-2027 CERTAIN

4.2.1 Domain Description

4.2.1.1 Digital Phenotyping for Mental Health Assessment

Digital phenotyping systems for youth mental health support early detection of anxiety and depression by identifying individuals at elevated risk based on multimodal behavioural and physiological markers collected during brief, standardized screening interactions. The domain is characterized by time-dependent, high-dimensional data streams, including audio-video recordings of tablet-based interviews, derived speech and language features, facial and movement cues, contactless cardio-respiratory measures, and environmental signals, all linked to validated symptom questionnaires and clinical information. From the perspective of clinicians and health systems, the primary objective is to improve timely identification and triage of mental distress while preserving privacy, which means that evaluation focuses not only on predictive accuracy but also on clinically meaningful outcomes such as sensitivity to elevated symptom burden, calibration of estimated risk, and robustness across diagnostic groups, age, and gender.

Within this context, the Syn.Mind cohort provides a clinically annotated benchmark for developing explainable models that map observable cues to anxiety and depression risk and for generating synthetic datasets that approximate the joint distributions and temporal dynamics of the underlying signals. Models range from traditional machine-learning approaches (for example logistic regression with feature selection) to more complex sequence models and generative techniques, and are assessed with respect to their ability to recover known associations between digital markers and symptom scores, maintain performance under class imbalance, and support privacy-preserving secondary use through synthetic data release rather than sharing of raw video or audio.

4.2.2 Contactless Physiological Signal Monitoring

Radiofrequency (RF)-based contactless vital signs monitoring is an emerging research field with clinical applications spanning several high-impact scenarios: monitoring of patients in ICU and burn units, where minimising physical contact reduces infection risk and patient discomfort; support for hospital-at-home and post-intervention recovery in independent living settings; extended clinical supervision for patients managing chronic conditions; seamless, unobtrusive physiological monitoring in home care or outpatient environments; and potential support for mental health research, as envisaged in CERTAIN WP7 (Health Pilot) on digital support systems for the early diagnosis and monitoring of childhood psychiatric disorders.

These scenarios are increasingly supported by AI-powered clinical decision support systems, capable of performing continuous, automated risk stratification, early deterioration detection, and alerting clinicians to abnormal physiological patterns, tasks where radar-based sensing offers a particularly compelling input modality. Its ability to operate without skin contact, across clothing, and independently of ambient lighting makes it well suited for longitudinal, low-burden monitoring in both clinical and domestic settings, where traditional wearable or contact-based sensors face compliance and usability challenges.

The methodology developed here focuses on the generation of synthetic breathing signals extracted from radar data, using apnea pattern and bad signal quality detection as a concrete and well-characterised problem to ground the approach. This choice is motivated by the availability of reference radar-based datasets covering these types of events, which provide a meaningful basis for validating both signal generation quality and the downstream AI models under test. Importantly, the scope of this work extends beyond this specific application: the objective is to establish a generalisable methodology for synthetic physiological time series generation that can be transferred to other radar-extracted signals and clinical tasks as relevant data become available, including the data to be collected within the WP7 Health Pilot.

4.2.3 Source Data Characteristics

Digital Phenotyping for Mental Health Assessment

Syn.Mind (clinical pre-screening cohort, being generated through CERTAIN project)

Syn.Mind comprises single-visit, tablet-based screening interviews with 12–20-year-olds, collected in a controlled hospital setting and now undergoing multimodal feature extraction and statistical analysis. Each visit links demographics and diagnostic grouping with standardized questionnaires (GAD-7, PHQ-9, BPFSC-11, LoPF-Q 12–18, BFI-2-S, ITAS, interview satisfaction, communication satisfaction), radar-based

© 2024-2027 CERTAIN

cardio-respiratory signals, indoor air-quality streams, and digital markers derived from synchronized audio–video of chatbot or virtual-agent interviews. The dataset consists of features of 252 participants.

Setting and population

Single-visit, observational pre-screening study in a hospital-based child and adolescent psychiatry setting at University Medical Centre Maribor (ages 12–20; clinical groups and controls).

Procedure

- Participants attend one in-person assessment in a dedicated study room, with standardized environmental preparation (air purification, sensor setup).
- They complete baseline questionnaires on a tablet, then a structured but conversational pre-screening interview delivered either by a text-based chatbot or a virtual agent, with random allocation to mode.
- The full interview is audio-video recorded; contactless radar (ViRa24) captures cardio-respiratory signals, and an indoor air-quality device continuously records CO_2, VOCs, temperature and humidity.
- After the interview, participants fill in post-interview satisfaction and communication scales, and clinician-reported diagnostic information is added where available.

SMILE (real-world resilience cohort, being generated in SMILE project, GA 101080923)

The SMILE project collects longitudinal data in naturalistic settings from 10–24-year-olds via a gamified digital platform and mobile tools, and is currently in the final stages of cohort recruitment and data collection. It includes repeated self-assessments of symptoms, resilience and social context, plus platform usage and engagement logs, supporting case studies across several European countries (United Kingdom, Germany, Cyprus, Poland, Slovenia, Spain, and Italy) in schools, universities and community settings. In total the data from 310 videos was used for CERTAIN project. The data is not directly ingested, but due to ethical concerns the features are ruin on separate pipelines and afterwards harmonized.

Setting and population

Multi-country, real-world cohort of young people (roughly 10–24 years) recruited via schools, universities, health and community services across several European countries (e.g. Slovenia, Italy, France, UK).

Procedure

- Participants register on a gamified digital platform (and associated mobile tools) designed to promote resilience and mental-health literacy.
- Over weeks to months, they complete repeated brief self-assessments (symptoms, resilience, social context), mini-questionnaires and diaries within the platform, alongside engagement with games, psychoeducational modules and challenges.
- The system logs usage and engagement traces (log-ins, time on modules, completion of activities, response to prompts) in naturalistic conditions, reflecting real-life patterns rather than controlled lab sessions

Relevant datapoints

Table 4-1: Relevant data points in the cohorts to be considered for the synthetic data generation

Modality / source	Datapoint examples (per participant or episode)
Demographics & context	Age, gender, education, country, setting.
Clinical status	Diagnostic group (anxiety, depression, BPD, control), current treatment, brief clinical history (Syn.Mind).
Interview meta	Interview mode (chatbot vs virtual agent), interview duration, language, session timestamp (Syn.Mind).
Symptom scales	GAD-7; PHQ-9 – or similar age relevant scores
Personality / traits	BPFSC-11, LoPF-Q 12–18, BFI-2-S item and scale scores; technology affinity (ITAS); resilience and connectedness scales (e.g. CYRM-12, BRS).

Draft

Facial affect & AUs	Percent of frames with basic emotions, lip press, down-angled mouth corners, smiling, frowns, anxious-face frames, emotion variability.
Gaze & head movement	Looking-down percentage, face-moving frames, head-down/facing-away (axes), head-moving frames, rotational energy (low/medium/high).
Speech prosody	Speech rate, articulation rate, number of pauses, speaking vs pausing duration, mean intensity, $f0$statistics, formant ranges, monotony indices.
Voice quality	Local jitter, shimmer (absolute/dB), harmonics-to-noise ratio and related noise/hoarseness measures.
Linguistic content	First-person pronoun use, sentiment proportions, negative vs positive word rates, explicit depression terms, absolutist words, past-focus, lexical diversity/density.
LLM text features	Counts/proportions of "death words", health and life-stressor words, negative/positive emotion words, basic text stats (sentence length distributions).
Radar physiology	Heart-rate and heart-rate-variability summaries, respiratory rate and variability, composite arousal indices per interview and per time window (Syn.Mind).
Air-quality environment	CO_2, VOCs, temperature, humidity: pre- and during-interview averages, variability, trends (Syn.Mind).

Contactless Physiological Signal Monitoring

Schellenberger et al., GUARDIAN Contactless Vitals Dataset (2020)[1]

This dataset was recorded using a 24 GHz continuous wave (CW) radar system and includes 30 healthy adult subjects (16 female), with an average age of 30.7 ± 9.9 years. The study was approved by the ethics committee of the Friedrich-Alexander-Universität Erlangen-Nürnberg (No. 85_15B). The informed consent was obtained from all subjects in human trials. Recording sessions covered four lie-down scenarios: a resting baseline of approximately 10 minutes, a tilt protocol (up and down, ~10 minutes), and two induced anomaly conditions: Valsalva maneuvers (3 × 20 seconds with 3 × 5-minute recovery periods) and apnea events (2 apnea episodes each followed by approximately 1 minute of recovery). Available data include raw radar signals alongside respiration, pulse, and heart sound waveforms, complemented by reference metrics. In total, the dataset spans more than 24 hours of recordings. Its combination of normal breathing periods and labelled anomaly segments makes it the primary source for learning both typical breathing morphology and the structure of clinically relevant anomalous patterns.

Yoo et al. Dataset (2021)[2]

This dataset was acquired using a 60 GHz frequency modulated continuous wave (FMCW) radar and covers 50 healthy children (26 female), ranging in age from 0.58 to 12.33 years. All the experiments were conducted in according to the guidelines provided by the Hanyang University local ethics committee (HYUH 2017-05-004). Prior to conducting the experiment, an informed consent signature was acquired from parents. For this task only children aged at least 60 months were selected, making the cohort of 23 healthy children (11 female). Each recording session consisted of approximately 5 minutes of resting, with available data including raw radar signals and estimated respiratory and heart rates. This is the only available radar dataset covering the paediatric population targeted by this use case. It enables the extraction of child-specific vital sign distributions

[1] https://www.nature.com/articles/s41597-020-00629-5

[2] https://www.mdpi.com/1424-8220/21/7/2412

© 2024-2027 CERTAIN

and radar signal parameters that are essential for generating age-appropriate synthetic data, ensuring that paediatric synthetic signals reflect realistic physiological characteristics rather than extrapolations from adult recordings.

4.2.4 Synthetic Data Generation Methodology

4.2.4.1 Digital Phenotyping for Mental Health Assessment

The synthetic data generation workflow was implemented as a multi-stage pipeline that transformed real-world data from the Syn.Mind and SMILE cohorts into privacy-preserving synthetic datasets for downstream modelling, validation, and benchmarking. Due to ethical and governance constraints, raw data from the two cohorts were not directly ingested into a single joint pipeline. Instead, Syn.Mind and SMILE data were processed in separate cohort-specific pipelines, and only de-identified, engineered, and governance-approved feature representations were harmonised at a later stage. This design preserved both cross-sectional feature relationships and, where relevant, temporal dynamics while respecting cohort-specific access restrictions and data-protection requirements. The methodology was designed to support explainable artificial intelligence, feature stability analysis, and privacy-aware reuse of clinically sensitive data.

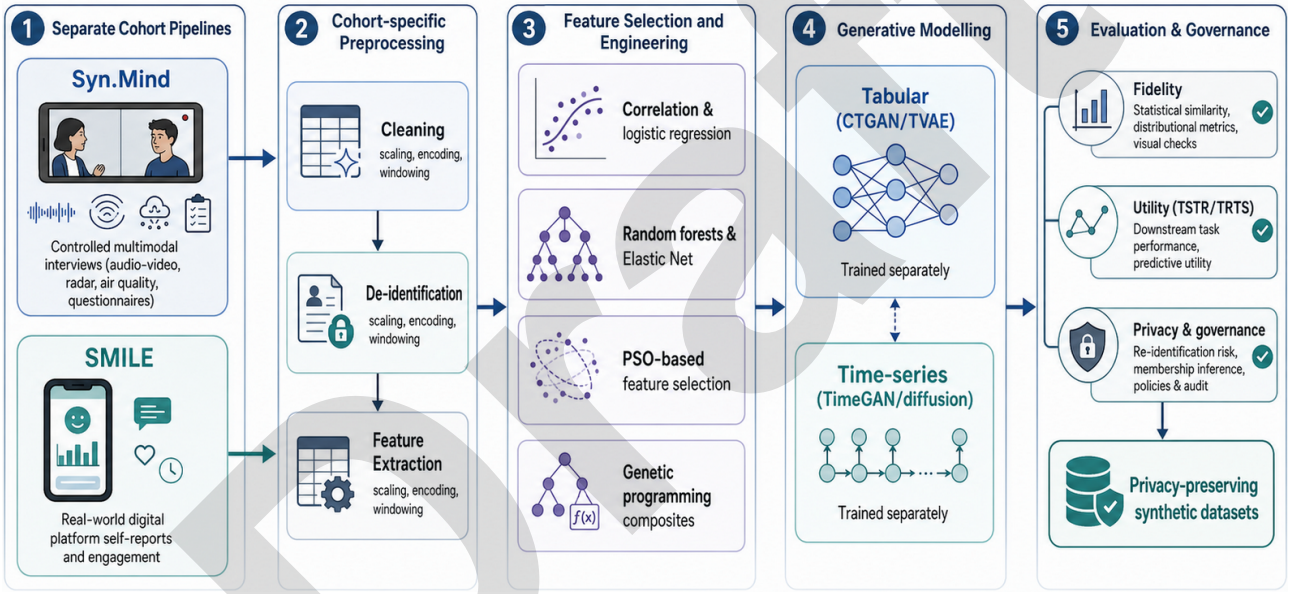

Figure 4-1: The CERTAIN Synthetic data generation & Evaluation pipeline. Note: This image was generated with the use of Chat GPT 5.4

Software stack

The implementation relied on the following Python ecosystem components:

- Data wrangling and numerical processing: pandas, numpy, scipy.
- Classical machine learning and preprocessing: scikit-learn, xgboost.
- Explainability: shap.
- Evolutionary and swarm optimisation: pyswarms, gplearn, deap.
- Deep learning and generative modelling: PyTorch, TensorFlow, sdv, ctgan, ydata-synthetic, and temporal generative frameworks such as TSGM.
- Visualisation and diagnostics: matplotlib, seaborn, plotly, and dimensionality reduction tools such as PCA, t-SNE, and UMAP implementations

Stage 1: Data preprocessing and harmonisation

© 2024-2027 CERTAIN

In the first stage, raw data from Syn.Mind and SMILE were processed independently within separate cohort-specific analytical pipelines rather than being directly pooled. Within each pipeline, episode-level tables were constructed for each interview, assessment, or digital interaction, and they were joined with questionnaire scores, demographics, contextual variables, and engineered observable cues. Only after local preprocessing, de-identification, and feature extraction were the resulting cohort-specific representations harmonised into a common analytical schema with aligned variable definitions, timestamps, cohort labels, modality tags, and outcome mappings. For selected modalities such as facial affect, speech prosody, radar-derived cardio-respiratory signals, and indoor air-quality measures, the raw or semi-processed streams were segmented into fixed-length temporal windows to form multivariate time series suitable for sequential generative models. This stage was implemented primarily in Python using pandas for structured data processing, numpy for numerical transformations, and scikit-learn preprocessing components such as SimpleImputer, StandardScaler, MinMaxScaler, and categorical encoders. Continuous variables were scaled either by z-score standardisation or min-max normalisation, categorical variables were encoded into model-compatible formats, and variable-length sequences were padded, truncated, or masked to a common representation prior to model training.

The second stage focused on identifying the most informative, stable, and interpretable features for synthetic modelling and downstream predictive utility. The process began with traditional statistical approaches, including descriptive summaries, correlation tables, and logistic regression models, in order to examine pairwise associations, detect multicollinearity, and identify candidate predictors of clinically relevant outcomes such as anxiety severity, depressive burden, resilience levels, or diagnostic group. Correlation matrices were used to screen for redundant variables and to distinguish potentially transdiagnostic cues from disorder-specific markers, while logistic regression served as an interpretable baseline for estimating direction and strength of association. These analyses were performed separately for tabular episode-level features and for summary descriptors extracted from temporal windows.

Stage 2: Identification of significant features

The traditional analyses were then complemented with baseline filter and wrapper methods using scikit-learn, including random forests, regularised logistic regression, Elastic Net, and recursive feature elimination with cross-validation, in order to rank variables according to predictive contribution and stability. This combination of classical statistics and machine learning supported an initial reduction of the feature space before more advanced optimisation was applied.

PSO was then used as a wrapper-based feature selection strategy to search the high-dimensional space of candidate feature subsets. Binary feature masks were optimized against composite objectives that combined predictive performance, subset sparsity, and feature stability across resampling runs. PSO was implemented using libraries such as pyswarms or custom optimization routines, and it was applied to downstream predictors such as logistic regression, random forests, or shallow neural networks to identify compact feature subsets with strong generalisation performance.

In parallel, GP was used for feature construction and symbolic combination of raw observable cues. Using libraries such as gplearn or deap, GP evolved interpretable composite indicators that integrated facial, prosodic, linguistic, physiological, and environmental variables into compact non-linear biomarkers. Candidate GP-derived features were evaluated according to predictive utility, parsimony, and interpretability, and only robust constructs with stable performance across folds and seeds were retained. To validate the final feature sets, SHAP-based explainability analysis was performed using the shap library, enabling comparison of feature contribution profiles across real-data models and later synthetic-data models.

Stage 3: Model selection and training

Tabular synthesis

For episode-level tabular data, the primary generative models included CTGAN and related synthesizers from the sdv ecosystem, as these models were specifically designed for mixed continuous-categorical data and conditional generation under class imbalance. Alternative tabular baselines included TVAE and, where appropriate, diffusion-based tabular generators implemented in PyTorch, especially for complex multimodal distributions with strong non-linear dependence structures. Model selection was guided by statistical fidelity, downstream task utility, and privacy metrics rather than by likelihood alone.

Time-series synthesis

For window-level multivariate temporal data, TimeGAN was used as a principal sequence generator, either through the original implementation or via wrappers such as ydata-synthetic. TimeGAN was used because it

© 2024-2027 CERTAIN

jointly learned latent representations, temporal transitions, and adversarial discrimination, which made it suitable for modelling behavioral and physiological dynamics. For more challenging sequential settings, temporal diffusion frameworks such as TSGM-based or PyTorch-based denoising diffusion architectures were also explored, particularly where longer-range dependencies or modality-specific temporal realism were required.

Hyperparameter optimisation

Hyperparameter optimisation was performed using a combination of conventional search and swarm-based optimisation. Grid search, random search, and optionally Bayesian optimisation tools such as optuna were used for baseline tuning, while PSO was also employed for black-box optimisation of generator hyperparameters such as latent dimensionality, generator and discriminator depth, learning rate, batch size, and training epochs. In selected experiments, GP was additionally used in a constrained architecture-search setting for smaller auxiliary models, although its primary role in the workflow remained interpretable feature construction rather than full neural architecture search.

Stage 4: Evaluation

The evaluation stage assessed synthetic data quality along three complementary dimensions: fidelity to the empirical data distribution, usefulness for downstream modelling, and privacy protection. Statistical fidelity was measured using univariate and multivariate similarity metrics such as Kolmogorov-Smirnov statistics, Jensen-Shannon divergence, pairwise correlation preservation, mutual-information similarity, and comparisons of principal-component embeddings between real and synthetic samples. For time-series data, additional metrics included autocorrelation preservation, cross-correlation preservation, Dynamic Time Warping distances, and visual comparison of trajectory families in reduced-dimensional spaces.

Utility was assessed through train-on-synthetic-test-on-real and train-on-real-test-on-synthetic protocols using predictive tasks such as anxiety prediction, depression prediction and diagnostic-group discrimination. Baseline predictive models for these utility experiments included logistic regression, random forests, gradient boosting, XGBoost, and GRU/LSTM sequence classifiers, implemented with scikit-learn, xgboost, and PyTorch or TensorFlow. Explainability consistency was also evaluated by comparing SHAP value distributions and feature-ranking concordance between real-trained and synthetic-trained predictive models.

4.2.4.2 Contactless Physiological Signal Monitoring

The synthetic dataset was developed through a four-step pipeline aimed at generating physiologically coherent breathing signals under controlled conditions. This approach combined data preprocessing and breathing signal extraction, the design of conditioning constraints, full time-series generation, and controlled pattern imputation. These steps ensured that the synthetic signals were not only realistic, but also generated in a way that reflected predefined constraints derived from the source data and the intended modelling objectives. The following steps describe how the synthetic dataset was generated, including the generation technique adopted and the constraints applied at each stage.

Step 1: Data preprocessing and breathing signal extraction

The three datasets described above were recorded with different radar technologies, frequency bands, and acquisition setups, making direct merging non-trivial. A unified preprocessing pipeline was applied to all recordings to extract and normalise the breathing signal from raw radar data, producing comparable time series across sources. This involved signal filtering, phase extraction or envelope detection as appropriate per radar type, amplitude normalisation, and annotation of available metadata (subject age group, session type, presence of artifacts or anomaly periods). The result is a single harmonised training corpus of breathing waveforms with consistent representation.

Step 2: Design of constraints for conditional generation

To ensure that synthetic signals are not only statistically plausible but clinically meaningful, generation is conditioned on two categories of constraints derived from external knowledge:

- **Age-dependent physiological constraints**: Breathing rate norm and breathing cycle morphology vary substantially between children and adults. Statistical parameters (e.g. respiratory rate distributions by age group) were extracted from the clinical and physiological literature and used to condition the generative models, ensuring that synthetic signals for paediatric subjects reflect age-appropriate physiology rather than extrapolating from adult data.

© 2024-2027 CERTAIN

- **Motion artifact constraints**: Noise patterns associated with patient movement were characterised as random noise with three distinct profiles. These are injected in a controlled manner during generation, producing signals with known, bounded noise characteristics suitable for robustness testing.

Step 3: Full time series generation

A range of machine learning generative architectures were implemented and evaluated, spanning from interpretable probabilistic models to deep generative networks, all adapted to the specific demands of physiological time series. As an interpretable reference point, a Gaussian Mixture Model combined with Principal Component Analysis (GMM-PCA) was included, offering a tractable generative process against which more complex approaches could be benchmarked. Moving towards deep learning, Variational Autoencoders (VAE) with GRU and 1D-CNN encoder-decoder variants were explored learning compact latent representations of breathing signal morphology and enabling principled sampling from a structured latent space to support conditional generation. Adversarial training was also explored through GRU-based Generative Adversarial Networks (GANs), where a generator and discriminator are trained jointly to increase the realism of the synthetic breathing signal. State-of-the-art methods (TimeVAE and TimeGAN) were included both as reference implementations reflecting current best practice and as foundations for custom architectural adaptations developed within this work.

Model evaluation was conducted using a complementary set of metrics targeting distinct aspects of synthetic signal quality:

- **Distributional fidelity** was assessed using *MMD (Maximum Mean Discrepancy)* and *Wasserstein distance*, computed both in a pooled setting (comparing the full synthetic and real distributions) and in a paired fashion (comparing matched segments), to measure how closely the generated signals replicate the statistical properties of real breathing waveforms.
- **Temporal structure** was evaluated using *autocorrelation analysis* and *temporal distance scores*, verifying that the periodic and sequential dependencies characteristic of breathing cycles are preserved in the synthetic signals.
- **Shape and distributional properties** were examined through *kurtosis* and *skewness* distributions, with pooled difference metrics used to quantify deviations from the reference signal morphology.
- **Latent space structure** was inspected via *t-SNE* visualisation, allowing qualitative assessment of whether synthetic samples occupy a realistic region of the data manifold and whether conditional generation (e.g. by age group) produces appropriately separated clusters.

Step 4: Controlled pattern imputation

Beyond full signal generation, targeted imputation was used to embed specific physiological and artifact patterns into otherwise normal synthetic signals. Apnea events were synthesised by imposing cessations of respiratory activity with duration and morphology constrained by clinically established parameters from the literature, and validated by extracting physiological features (e.g. apnea duration) and comparing them against published reference ranges. Motion noise segments were similarly imputed using parameterised artifact templates. This approach enables precise control over the frequency, duration, and severity of both anomaly types in the final synthetic dataset.

4.2.5 Intended Use for Compliance Testing

4.2.5.1 Digital Phenotyping for Mental Health Assessment

The generated synthetic data are intended to support conformity-oriented technical validation of AI-based youth mental-health screening and risk-assessment systems within the CERTAIN health pilot. The synthetic data contribute to the assessment of compliance-relevant characteristics required under the EU AI regulatory framework, including accuracy, robustness, fairness-related performance, and transparency of the AI system.

Transparency, auditability, and traceability: by enabling systematic testing of AI system behaviour on realistic but artificial inputs, in line with CERTAIN objective O1 on traceability and O4 on methods to assess compliance of AI systems.

Fairness and bias assessment: including whether model performance varies across demographic and clinical subgroups, consistent with the CERTAIN health pilot evaluation purpose, which explicitly includes

© 2024-2027 CERTAIN

fairness assessment, evaluation of training data for biases, and monitoring of performance across demographics.

Evaluation of robustness: by stress-testing the AI system under controlled and realistic scenarios, as required by CERTAIN objective O4, which targets technical requirements such as fairness, robustness, and transparency through generated test data.

Privacy preservation and data minimisation: enabling testing and validation activities to be conducted on synthetic rather than directly identifiable psychiatric data, which is aligned with the Syn.Mind protocol's privacy-by-design approach and CERTAIN's emphasis on privacy-preserving synthetic data generation.

Clinical and technical validity: by checking whether synthetic samples preserve the relevant statistical structure, multimodal relationships, and disorder-related feature patterns required for meaningful testing of digital mental-health AI systems.

Regulatory compliance and certification-oriented evaluation: quantified evidence supporting CERTAIN's broader objective of equipping AI system providers and certification entities with methods, tools, and test evidence for conformity assessment under the EU AI Act and related EU data-regulation frameworks.

4.2.5.2 Contactless Physiological Signal Monitoring

For the health domain, and particularly with regards to the radar-extracted breathing signals, this task produced two different testbed datasets, aimed at two compliance testing objectives: Fairness and Robustness:

Fairness: Data includes synthetic breathing signals generated for both adult and paediatric populations, with age-conditioned generation ensuring that physiological characteristics are realistic for each group. This enables evaluation of whether AI-powered monitoring models produce consistent and unbiased outputs across age groups, a dimension of fairness particularly relevant under the EU AI Act's requirements for high-risk AI systems in healthcare, where systematic performance disparities across demographic subgroups must be identified and mitigated.

Robustness: Data includes synthetic breathing signals with controlled inclusion of motion artifact patterns for systematic evaluation of model stability under realistic noise conditions. Specifically, it is possible to assess whether signal quality detection components correctly identify motion-corrupted segments, and whether apnea detection performance degrades gracefully in the presence of such artifacts. The ability to modulate noise intensity and duration provides a structured framework for stress-testing models along a realistic operating envelope.

4.2.6 FAIR Principles for Data Sharing

4.2.6.1 Digital Phenotyping for Mental Health Assessment

All curated synthetic datasets and source code openly available under a Creative Commons Attribution (CC BY) licence, subject to applicable ethical and legal constraints. Version-controlled code, analysis pipelines, and documentation will be maintained in a GitLab repositories: https://gitlab.com/humadex/mental-health-analysis-framework and https://gitlab.com/humadex/mental-health-synthetic-data-framework (to be finalized and publicly open by M30 of the project when they have been evaluated in the targeted clinical setting), ensuring transparent development history, issue tracking, and community contributions, while stable releases of code and data will be archived in trusted repositories that assign persistent identifiers (e.g. DOIs).

In line with the FAIR (Findable, Accessible, Interoperable, Reusable) data principles, all shared artefacts will be accompanied by rich, standards-compliant metadata, clear descriptions of provenance and processing steps, and machine- and human-readable licensing information. Findability and accessibility will be supported through deposition in repositories that expose searchable metadata and open access via standard protocols, while interoperability will be promoted through the use of non-proprietary formats and domain-relevant controlled vocabularies. Reusability will be further strengthened by providing detailed data dictionaries, codebooks, and example analysis notebooks, and by preparing a peer-reviewed data descriptor to be submitted to Scientific Data (Nature Portfolio) or a comparable high-impact journal specialising in FAIR-aligned data publications.

Draft

4.2.6.2 Contactless Physiological Signal Monitoring

All synthetic datasets produced within this task are openly released alongside the complete codebase, including model implementations, configuration files, and experiment scripts, through a public GitHub repository under a Creative Commons Attribution (CC BY) licence to ensure that the experimental pipeline can be re-executed in an identical software environment.

- code: https://github.com/CMazzolaNVISION/CERTAIN_T5_1_Health_Radar.git
- datasets: https://github.com/CMazzolaNVISION/CERTAIN_T5_1_Breathing_Testbeds.git

The datasets are distributed in CSV format, chosen for its compatibility with the wide ecosystem of machine learning and data analysis tools commonly used in the research community. To allow independent reuse and interpretability, the repository provides thorough documentation covering feature definitions, metadata structure, and annotation conventions.

These choices reflect a commitment to the FAIR principles throughout the data lifecycle. The public repository ensures that datasets and code are findable and immediately accessible. Standardised formatting and explicit schema documentation support interoperability across different tools and research contexts. Comprehensive documentation combined with versioned, reproducible code maximises reusability, allowing other researchers and practitioners to build on, validate, or extend this work with confidence.

4.3 BIOMETRICS

Biometrics is the only domain the deliverable covers from two perspectives. The domain itself, face recognition for identity verification, is shared and described once below; what differs is who runs the tests and what the synthetic data must provide. A system provider (IPS) tests its own identity-verification systems, so it generates matched reference and live face images that exercise the enrolment and verification pipeline at volumes real image acquisition cannot reach. An independent test lab (Idiap) evaluates face recognition systems it did not build. Its datasets must vary in a controlled way across capture conditions, identity-bound attributes and soft-biometric factors. That variation lets the lab measure performance and bias and audit identity leakage for any system under test. Both perspectives rely on generative face models. The requirements on the resulting datasets differ with the role, and the two subsections document them separately.

4.3.1 Domain Description

IPS explored synthetic image generation to evaluate ID-verification systems that employ facial recognition technology (FR). Such systems operate in two stages: 1) enrolment; 2) verification.

Enrolment: During the enrolment stage, a high-quality reference image is captured under controlled circumstances (frontal; neutral illumination; neutral expression). Optionally, a template of biometric features may then be extracted from the image and either stored in a secured database or securely embedded in an ID document. Alternatively, the image itself may be embedded in the ID document for extraction of the reference template immediately prior to verification.

Verification: During verification, a live image is captured at the secure entry point and is encoded by the same algorithm as used to process the reference image. Matching of templates is then performed, typically by calculating the cosine-similarity between the reference and live templates. The resulting similarity score is then compared to some pre-defined threshold to decide whether access is allowed or not.

The following sections outline a method for generation of synthetic reference and live images representing subjects that have been imagined by a generative adversarial network (GAN) and further developed using a diffusion model (DM). We show that our generated dataset has been anonymised with respect to the various training datasets and that it can be used successfully to measure trends in algorithmic performance.

Draft

© 2024-2027 CERTAIN

4.3.2 Usage of synthetic data by system provider in Biometrics

4.3.2.1 Source Data Characteristics

The datasets produced using our proposed generation method benefit from the various characteristics of a number of FR datasets. Details of these datasets are given in Table 4-2.

Table 4-2 Datasets of bona fide face images used for training our generative models.

Dataset	Type	Num. images
WebFace21M Restored [9]	In the wild (tight crop; variable quality; restored using GFPGAN [10])	21M
FFHQ [11]	In the wild (high quality; relatively loose crop)	70k
CelebA-HQ [12]	In the wild (high quality; relatively loose crop; restored using jpeg removal and super-resolution [13]	30k
Operational Refs	Controlled reference images (good quality; various crops)	41k

The various datasets described in Table 4-2 contribute to our synthetic datasets in complementary ways: WebFace21M enables pre-training for a vast dataset of many identities and images from which a strong identity prior can be learned. Image quality in WebFace21M, however, is low, and so the smaller but higher quality datasets – FFHQ and CelebA-HQ – are used to learn a high-quality, in-the-wild image-style. Finally, an internal dataset of operational reference images is used to adapt from in-the-wild to controlled-style images.

4.3.2.2 Synthetic Data Generation Methodology

To generate diverse but identity-consistent sets of synthetic images, we rely on the success of biometric-template-conditioned diffusion models; namely, Arc2Face [16]. Arc2Face takes an ArcFace [17] biometric template as conditioning input and, by employing classifier-free guidance [18], generates a corresponding face-image from random noise via iterative denoising of a noise vector in the latent-space of a variational autoencoder (VAE) [19]. Arc2Face is a fine-tuning of StableDiffusion [20] and is mainly trained using the WebFace21M dataset. Due to the large number of images and the diversity of identity in the dataset, the model generalises well and is able to interpolate to generate mated sets of images for identities that are not present in the training datasets. In our initial attempts at sampling new identities, we sampled random vectors from a standard Gaussian distribution before normalising to give a random point on ArcFace's hyperspheric identity manifold. Unfortunately, doing so resulted in a high proportion of "unreal"-type images being generated, i.e. drawings and paintings. These identities may have been learned from fictional characters or from historical figures for which no photos exist. To avoid this issue, we first generated synthetic identities using a GAN trained only on real subjects.

GAN identity sampling

Other works have used GANs to first perform synthetic identity sampling prior to generation of mated image sets [21]. Whereas [21] optimises the distribution of points in the GAN's latent space to produce evenly distributed identities on the surface of the biometric hypersphere, therefore optimising inter-class distance, we, instead, take a simpler, brute-force approach: we first randomly sample 1M random points in the GAN's latent space to generate 1M synthetic images and then filter the images as follows:

1. We remove all images of insufficient quality based on an internal universal biometric quality metric.

© 2024-2027 CERTAIN

2. We remove all images of subjects estimated to be younger than 12 years old, since facial recognition networks are known to function less well for children. Allowing images of children to remain would cast doubt as to whether a synthetic dataset is truly anonymous or not.
3. We remove images to ensure that no two synthetic images give a cosine-similarity of greater than 0.4 in the hyperspheric feature space of our FR network, "IPS_2025". In practice, rather than optimising this selection, we take a greedy approach and select random images sequentially, only adding them to the set-to-keep if the image matches all other images currently in the set with a sufficiently weak cosine similarity, i.e.< 0.4.
4. Finally, we estimate the gender and skintone of the selected images and remove images to roughly balance groups as shown in Table 4-3.

Table 4-3 Synthetic identity counts per demographic group. Our skintone scale spans from 0.0 (light skin) to 1.0 (dark skin).

Skintone	Male	Female	Total
0.0-0.2	1287	1298	2585
0.2-0.4	1275	1269	2544
0.4-0.6	1351	1353	2704
0.6-0.8	1306	1328	2634
0.8-1.0	1247	1231	2478
Total	6466	6479	**12945**

As in [21], StyleGAN2 [5] was used to sample synthetic identities. We opted to use our own version of StyleGAN2 retrained on a 128x128 version of FFHQ. We found that this lower-resolution GAN was preferable to the official, 1024x1024-resolution version of the GAN as it gave lower false acceptance rates for random distributions of generated images. We believe that, in our lower resolution version of the GAN, less network capacity is wasted on biometrically irrelevant, high-resolution detail, allowing a larger section of the distribution of identity to be captured. Note also that truncation of the latent vector was disabled when generating our initial set of 1M synthetic images. Examples of the GAN-generated images can be seen in the first column of Figure 4-3.

ID-conditioned Diffusion Model

For each of the GAN-generated, synthetic identities, we wish to generate mated sets of high-resolution, high-quality images. For this, we choose to use a template-conditioned diffusion model; namely Arc2Face [16]. Arc2Face is conditioned on ArcFace templates, a method which dates from 2018. Figure 4-2 shows that a more recent FR algorithm, IPS_2025, significantly outperforms ArcFace. To benefit from the higher performance seen in Figure 4-2, we performed our own fine-tuning of StableDiffusion, replacing text-conditioning with template-conditioning as outlined in [16]. We refer to this model as IPS2Face.

Draft

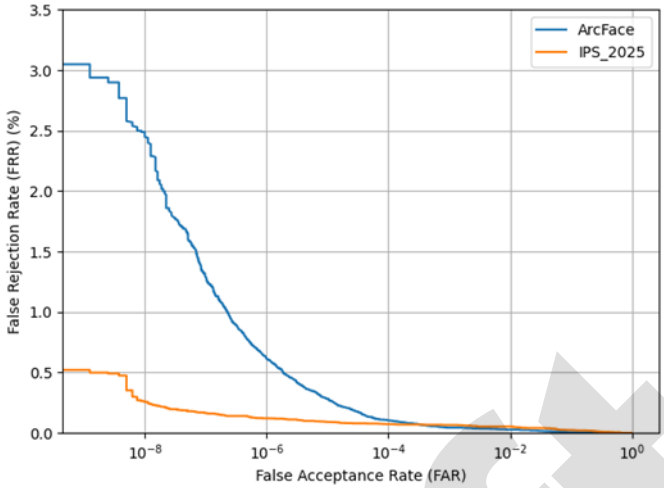

Figure 4-2 DET curves comparing the performance of ArcFace and IPS_2025 for a dataset of real images.

IPS2Face was first fine-tuned on WebFace21M followed by further fine-tuning on FFHQ and CelebA-HQ. We also trained reference image versions of both Arc2Face and IPS2Face by further fine-tuning the two models on our internal dataset of operational reference images. We will refer to these versions of the networks as *Arc2Face-ref* and *IPS2Face-ref*. Examples of the images generated by IPS's various generative networks can be seen in Figure 4-3.

Draft

© 2024-2027 CERTAIN



Figure 4-3 Examples of synthetic images generated by our 128x128-resolution StyleGAN2 (column 1), by IPS2Face-ref (column 2), and by IPS2Face (columns 3, 4, and 5).

In Figure 4-4 we show matching performance for the MagFace [22] facial recognition algorithm applied to synthetic datasets generated using Arc2Face, IPS2Face, and also the combinations Arc2Face+Arc2Face-ref and IPS2Face+IPS2Face-ref that we will refer to as Arc2Face-live+ref and IPS2Face-live+ref respectively. For the combined datasets, matching scores are always bipartite between the sets of reference and live images to reflect an operational scenario. MagFace was chosen for this evaluation as it is independent from both ArcFace and IPS_2025. During evaluation of our synthetic datasets, we noticed that their biometric consistency could be over-estimated if using a different but related network, for example, when evaluating IPS2Face using another IPS network that shares the same architecture or training procedure. This is presumably due to any biases of the network being reflected in the generated images. From Figure 4-4 we see that for both Arc2Face and IPS2Face, addition of the reference images reduces false rejection rates as intended. Overall, the best identity-consistency is measured for IPS2Face-live+ref, at least at more demanding operating points.

Co-funded by the European Union
© 2024-2027 CERTAIN

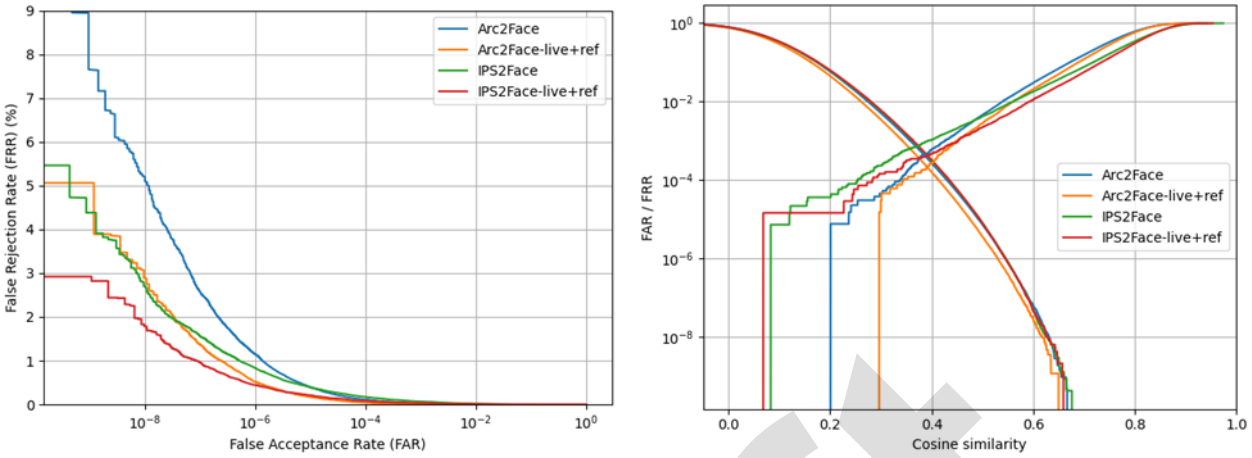

Figure 4-4 MagFace matching statistics for synthetic datasets generated using variants of Arc2Face and IPS2Face.

Note that stronger mated matching scores observed for IPS2Face could simply be due to limited intra-class variation in images, i.e. less variation of pose, illumination, expression, age, etc. This, however, is not the case as can be seen in Figure 4-5 which shows that LPIPS [23] distances within mated sets of images are actually larger for IPS2Face. This indicates a higher level of perceptual variation within mated image sets despite variation in identity being better constrained. It stands to reason that a biometric template that is less correlated with non-identity factors would allow those factors to vary more freely during generation.

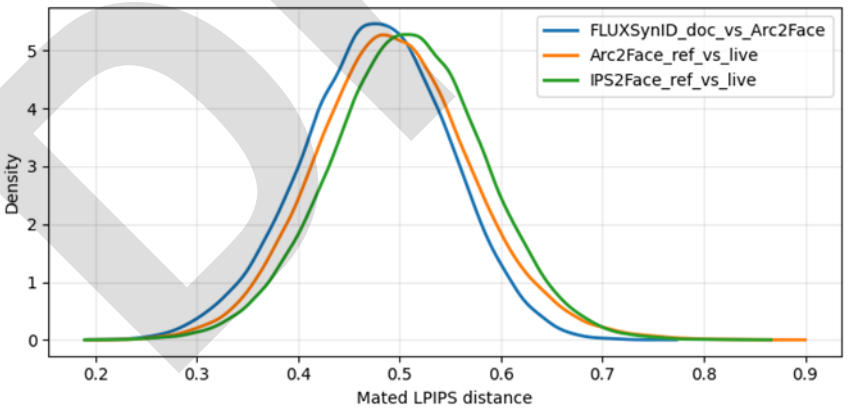

Figure 4-5 Distributions of LPIPS distances for three synthetic datasets revealing the overall perceptual variety within mated sets of images.

Comparison with FLUXSynID

FLUXSynID [24] is arguably one of the highest quality synthetic FR datasets currently available publicly. It contains 14889 synthetic reference images generated by the FLUX.1 model [25]. Demographics and other characteristics were balanced via appropriate text-conditioning. Live images were then generated from the FLUX.1 refs using Arc2Face. Figure 4-6 shows that the false acceptance rates for our GAN-based set of identities are slightly lower than those of FLUXSynID. Indeed, whereas we ensured that none of our synthetic identities strongly matched any other, FLUXSynID allowed for an expected rate of biometric similarity for the size of their dataset. In Figure 4-6, we also see that the genuine curve (the false rejection rates) shows a

Draft

© 2024-2027 CERTAIN

different dynamic with a higher proportion of strongly matching reference-live mated pairs. This may be due to the fact that FLUXSynID used reference images to condition Arc2Face directly whereas our reference and live images are linked only indirectly via the templates extracted from our GAN-generated images. We also observe in Figure 4-5 that there is less perceptual variation between FLUXSynID's reference and live images compared to that for our datasets. Again, this could be due to the live images directly inheriting characteristics from the reference image due to lack of robustness of the ArcFace templates to non-identity factors.

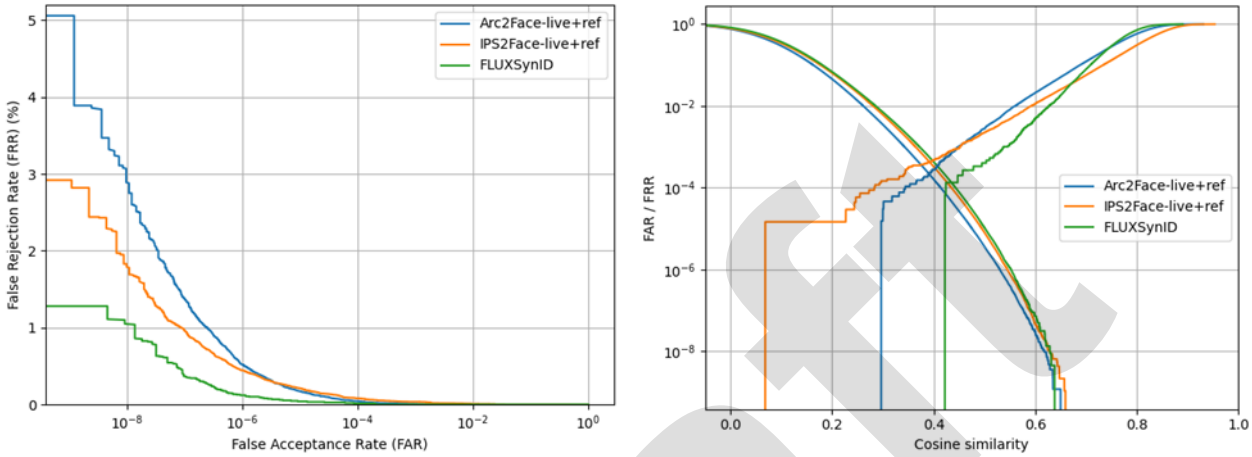

Figure 4-6 MagFace matching statistics for synthetic datasets Arc2Face-live+ref, IPS2Face-live+ref and FLUXSynID.

Arc2Face synthetic dataset generation

For each of the datasets evaluated above, and in the applications of the following section, five live images were generated per identity. For Arc2Face-live+ref and IPS2Face-live+ref, an additional reference image was generated by first generating five reference images and then selecting the best in terms of measured biometric quality. Note that, since the diffusion models and their fine-tuned, reference-image-generating counterparts are strongly related, to avoid similarities, it was necessary to use different sets of seeds for generation of the reference and live images. Note also that, unlike [21], no perturbing of the ID vector was used to increase intra-class variation, which can be done only at the expense of identity-consistency.

Despite having been able to improve the variety and consistency of images, in the following section we choose to evaluate our FR algorithms using *Arc2Face*-live+ref. Doing so avoids the issue of algorithmic shortcomings being masked by compensating errors in the synthetic dataset.

Verification of dataset anonymity

Before progressing to operational use of Arc2Face-live+ref, we first perform final checks to ensure that the synthetic dataset does not leak biometric information from the various training datasets. To do so we match all synthetic images against all images from the various training datasets. In Figure 4-7 (left) we show FAR curves for the Arc2Face live images matched against the training datasets WebFace21M, FFHQ and CelebA-HQ. We also show the FAR curve for impostors within a clean dataset of real, in-the-wild images. In Figure 4-7 (right) we show the FAR curve for the Arc2Face-ref images matched against our operational reference image dataset and also the FAR curve for the real reference images matched amongst themselves. In both Figure 4-7 (left) and (right) we see that the synthetic-to-real false acceptance rates are lower than those seen for the real images themselves. This means that, even though synthetic images may match training images more or less strongly, the rate of matches is always lower than that of similarly strong matches observed in a random population. The dataset as a whole is therefore anonymised, and there is no evidence of systematic re-generation of training identities.

Draft

© 2024-2027 CERTAIN

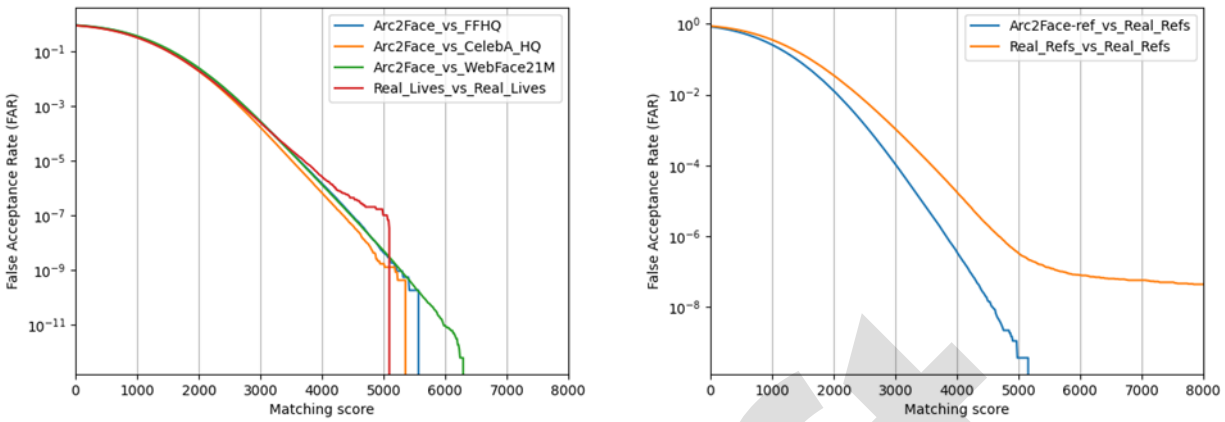

Figure 4-7 False acceptance rates for synthetic-to-real impostor pairs and real-to-real impostor pairs. Synthetic images were generated using StyleGAN2 + Arc2Face (left) and StyleGAN2 + Arc2Face-ref (right). The FR algorithm IPS_2025 was used for coding and matching.

4.3.2.3 Intended Use for Compliance Testing

Although evaluation using synthetic datasets does not always produce the same performance statistics in absolute terms, the datasets can nevertheless be used successfully to make comparisons. In Figure 4-8 we compare the performance of three IPS algorithms using an operational dataset of real reference/live images and our Arc2Face-live+ref synthetic dataset. Ignoring the change in scale on the vertical axis, we see that the DET curves for the two datasets show very similar characteristics. This result shows how mated sets of synthetic images might be used to ensure the successful update of an FR system at a client site.

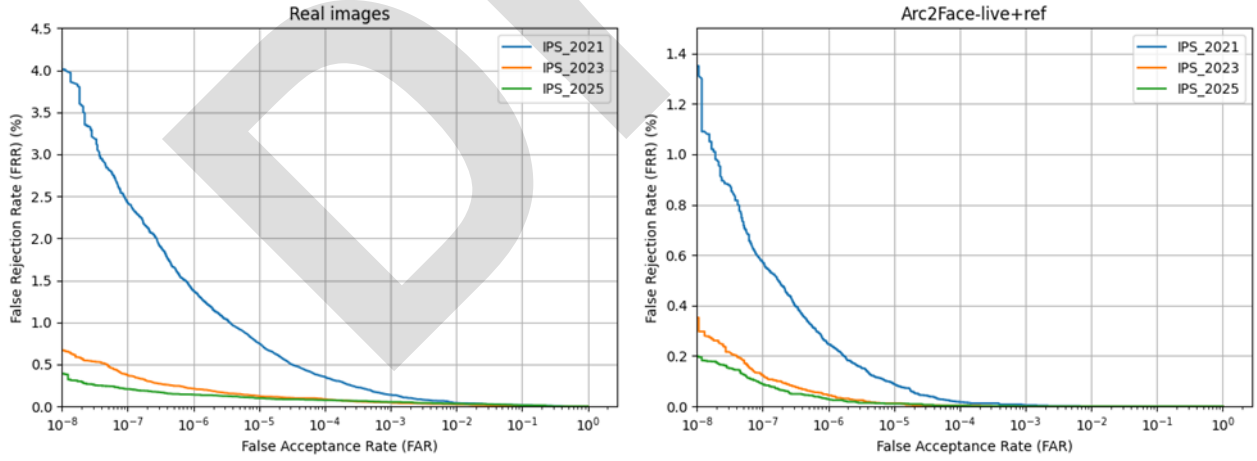

Figure 4-8 Performance statistics for various IPS FR algorithms evaluated using an operational dataset of real images (top) and our synthetic Arc2Face-live+ref dataset (bottom).

In Figure 4-9 we see that appropriate subsets of the synthetic datasets can be used to measure demographic bias. We show performance for two IPS FR algorithms applied to real and synthetic images measured as having skintone levels 0.0-0.2 (“Light skin) and 0.8-1.0 (“Dark skin”). Again, we see similar qualitative characteristics between the curves for real and synthetic data: in both datasets, we observe a bias for the algorithm IPS_2021, in which light skinned identities suffer from a higher level of false rejection. Between the algorithms IPS_2021 and IPS_2025, we then see a general improvement of the algorithmic robustness, and also a reduction in the level of bias. Again, this result shows how mated sets of synthetic

Draft

© 2024-2027 CERTAIN

images might be used to ensure the successful update of an FR system where the update might target reduction of bias.

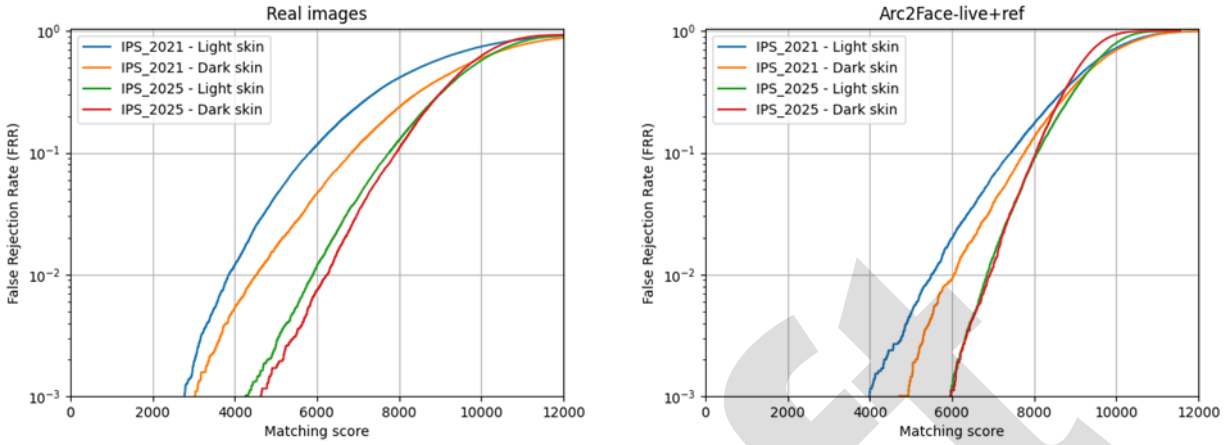

Figure 4-9 Performance statistics for two IPS FR algorithms evaluated on subsets of varying skintone for both an operational dataset of real images (left) and our synthetic Arc2Face-live+ref dataset (right).

4.3.3 Usage of synthetic data by test lab in Biometrics

4.3.3.1 Source Data Characteristics

Methods suitable for testing labs draw on a large number of named datasets, which fall into three functional categories. These are (i) *generator training corpora* used to train or fine-tune the generative models, (ii) *evaluation benchmarks* on which downstream face-recognition accuracy is measured, and (iii) *auxiliary datasets* used for identity sampling, covariate direction learning, or on-manifold regularisation. The key statistics for all datasets are collected in tables further into the section.

Furthermore, Idiap considered the following characteristics for the evaluation of generated synthetic datasets:

- **Capture (extrinsic) characteristics.** These arise from the acquisition conditions rather than the subject, and can in principle be constrained by the capture setup. They include head *pose* (yaw, pitch and roll), *illumination* (direction, intensity and colour temperature), facial *expression*, image *quality and resolution*, and *background* and sensor properties.
- **Identity-bound (intrinsic) characteristics.** These are properties of the subject that cannot be altered by the capture setup. They include *age*, *skin tone*, *gender* and *ethnicity* or ancestry. Because they are bound to the individual, a deployed system cannot constrain them. It must accept whoever presents themselves, and its behaviour across the resulting sub-populations is fixed at training time.
- **Soft-biometric characteristics.** These are attributes that carry partial identity information and sit between the two groups above. They include *hair* (style, length and colour), *facial hair*, *eyewear* such as glasses and sunglasses, *makeup*, and other *accessories*. Some are transient and partly controllable, while others are stable over the short term.

Table 4-4 Datasets as part of the training corpora for synthetic data generation methods considered by Idiap.

Dataset	Identities	Num Images	Collection method	Used for
WebFace42M (subset) [26]	1M	21M	Web-crawl + auto-clean	Arc2Face generator training

Draft

WebFace4M (subsets) [26]	50K–200K	1M–4M	Web-crawl + auto-clean	Vec2Face generator training
FFHQ [23]	N/A	70K	Flickr scrape, diverse	StyleGAN2 (DisCo/HyperFace), Arc2Face stage-2 FT
CelebA-HQ [22]	N/A	30K	Celebrity images, high-res	Arc2Face stage-2 FT
MS-Celeb-1M / MS1MV2 [27] [28]	85K–100K	5.8M–10M	Web celebrity scrape	DisCo identity backbone, Vec2Face PCA sampling
Glint360K [29]	360K	17M	Web-crawl + auto-clean	Vec2Face feature extractor training
CMU Multi-PIE [30]	337	750K	Controlled studio	DisCo covariate directions (pose/illum./expr.)
CASIA-WebFace [31]	10K	494K	Web-crawl	Real-data 0.5M-scale reference baseline
BUPT-Balancedface [32]	28K	1.3M	Web-crawl, race-balanced	HyperFace on-manifold regularisation gallery

The standard practice in synthetic face-recognition research to benchmark the quality of a synthetic dataset is to train a face-recognition backbone on the synthetic dataset under evaluation and then report verification accuracy on a fixed set of public benchmarks. The main motivation behind this benchmark is to quantify how much information, variation, quality, and fidelity are maintained by a generated dataset in contrast to datasets of real face images. If the performance of a face recognition model trained on synthetic data is close to one that was trained on real data, then one can plausibly claim that the synthetic dataset contains as much information that can be extracted by a deep neural network-based face recognition as a real dataset. Therefore, we report the face recognition performance of face recognition models trained on synthetic data evaluated on real data as a proxy for the quality of the synthetic dataset.

Relevant benchmarking datasets with real face images are summarised in Table 4-5 below. A subset of these benchmarks is used to evaluate the quality of the synthetic data generation methods considered and/or proposed by Idiap in the following sections.

Table 4-5 Datasets as part of the evaluation of synthetic data generation methods considered by Idiap.

Benchmark	Subjects	Pairs / templates	Test axis	Metric
LFW [33]	5,749	6,000 pairs	Unconstrained, near-frontal	TMR @ FMR
CFP-FP [34]	500	7,000 pairs	Frontal vs. profile	TMR @ FMR
CPLFW [35]	3,000	6,000 pairs	Cross-pose (LFW repairing)	TMR @ FMR
AgeDB-30 [36]	568	6,000 pairs	~30-yr age gap	TMR @ FMR
CALFW [37]	3,000	6,000 pairs	Cross-age (LFW repairing)	TMR @ FMR
IJB-B [38]	1,845	12,115 templates	Unconstrained media-set	TMR @ FMR (1e-4)
IJB-C [39]	3,531	23,124 templates	Unconstrained media-set	TMR @ FMR (1e-4)

Draft

© 2024-2027 CERTAIN

RFW [32]	~3K/group	~6Kx4 pairs groups	Racial bias	Fairness metrics
Twins-IND [40]	219	6,000 pairs	Near-identical (twins)	Verification accuracy
DoppelVer [41]	390	doppelgänger pairs	Near-identical (non-twins)	Verification accuracy
Hadrian	2,600	6,000 pairs	Facial-hair variation	FID / verification accuracy

Synthetic Data Generation Methodology

We considered two methods for generating synthetic data that are identity preserving and have sufficient intra-identity variability in terms of the source data characteristics discussed in 4.2.3. The details of the methods, namely DisCo and HyperFace, and their relative performances when used to train a face recognition are given in the rest of this section.

The code, pretrained models, project pages and source publications that make the two reviewed methods accessible and reusable are collected below.

- **DisCo** at https://www.idiap.ch/paper/synthetics-disco
- **HyperFace** at https://www.idiap.ch/paper/hyperface

a. DisCo

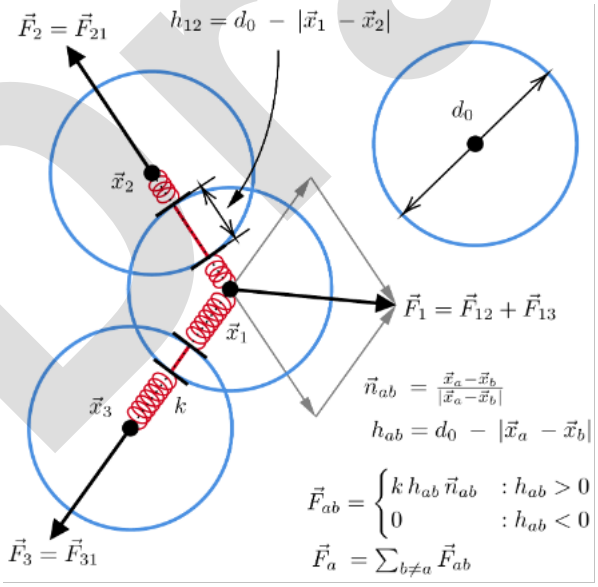

Figure 4-10 Soft-particle contact-force model underlying the DisCo Langevin and Dispersion dynamics, reproduced from the paper "Synthetic Face Datasets Generation via Latent Space Exploration from Brownian Identity Diffusion"

DisCo considers synthetic-identity generation as the stochastic motion of soft spherical particles in StyleGAN2's W-space, formalised via the over-damped Langevin SDE from statistical mechanics. All three algorithms (Langevin, Dispersion and DisCo) share the same numerical recipe, gradient descent on a granular-mechanics loss: quadratic repulsion when particles overlap, plus quadratic pull-back to w_{avg} (the mean of the StyleGAN2 W-space distribution, used as a quality anchor for the optimisation, augmented with a Gaussian thermal noise term. Langevin spreads identities in the face-recognition embedding space, where the granular

Draft

force acts in E, over approximately 100 iterations. Dispersion generates intra-class variations of each identity with the granular force acting in W instead, together with a spring attractor in E that holds the embeddings near the anchor, and its initialisation is small Gaussian noise. DisCo differs from Dispersion only in its initialisation, which is wider. The start is offset by a random linear combination of the seven pre-computed Multi-PIE-derived covariate directions in W, so the optimiser's starting point already spans pose, expression and illumination. The end-to-end synthetic dataset pipeline is therefore Langevin (for inter-class spread) followed by DisCo (for intra-class spread).

DisCo is part of a unified three-algorithm framework that separates the inter-class problem (different identities should be different from each other) from the intra-class problem (each identity should appear under varied attributes). Langevin handles inter-class spread. It iteratively pushes a set of identity-anchor latents apart in the face-recognition embedding space using a granular-mechanics repulsive force, together with the quadratic pull-back and a stochastic Brownian force. Dispersion handles intra-class diversity by applying the same Brownian and granular machinery around each anchor, although the repulsion now acts in StyleGAN2's W-space (pushing intra-class samples apart so the images differ) while a spring attractor in embedding space keeps the synthetic identities close to the anchor. DisCo, an acronym for Dispersion combined with Covariates, modifies the Dispersion initialisation. Each intra-class variation's starting latent is offset by a random linear combination of the seven Multi-PIE covariate directions, with weights uniformly sampled in (default $[-\lambda_0, \lambda_0]$). Starting from this wider initialisation, the optimiser explores a larger intra-class manifold along pose, illumination and expression. DisCo gains roughly 12 percentage points over plain Dispersion on the pose-heavy benchmarks CFP-FP and CPLFW. Released datasets reach 30K identities and 1.95M images.

At 10K identities and 0.65M images (65 images per identity), Langevin+DisCo reaches a 5-benchmark average of 85.16% (LFW 97.07, CFP-FP 79.56, CPLFW 76.73, AgeDB 83.38, CALFW 89.05). At 30K identities and 1.95M images the average rises to 90.31%. Standard image-quality metrics such as FID or KID (Kernel Inception Distance, an unbiased estimator of distributional distance between generated and real images, similar in spirit to FID) are not reported. Image quality is discussed only qualitatively through the pull-back mechanism that keeps generated images near StyleGAN2's high-quality region. The framework reports the lowest RFW racial-bias standard deviation (1.81 percentage points) among the synthetic datasets it compares against, below DCFace (2.66), IDiff-Face (2.72) and GANDiffFace (3.00).

b. HyperFace

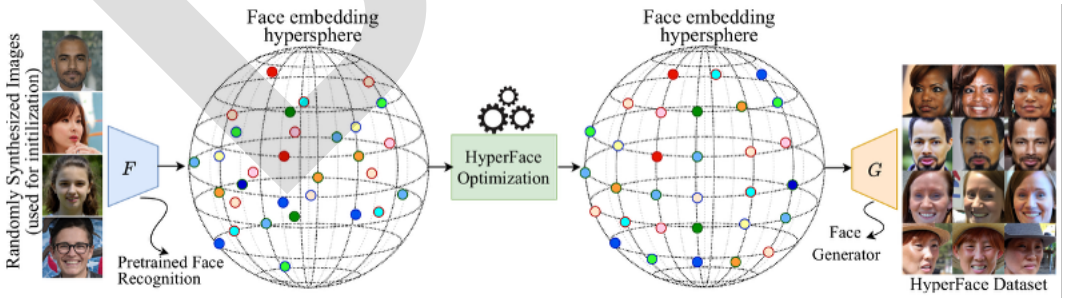

Figure 4-11 Block diagram of the HyperFace dataset-generation pipeline, reproduced from the paper "HyperFace: Generating Synthetic Face Recognition Datasets by Exploring Face Embedding Hypersphere" [16].

HyperFace formulates synthetic dataset generation as a spherical-code packing problem on the 512-d ArcFace hypersphere, where the goal is to find unit-norm reference embeddings that maximise the minimum pairwise angular distance. The objective combines two terms. The first is a max-min loss that drives each pair of references to its angular-distance ceiling. The second is a regularisation term that pulls each reference toward its nearest neighbour in a gallery of real or synthetic-face ArcFace embeddings, which prevents the optimiser from drifting off the face-embedding manifold and producing identities the image generator cannot render (optimisation weight 0.5 by default). The optimisation variable is the set of reference embeddings, initialised

© 2024-2027 CERTAIN

from ArcFace features of StyleGAN-synthesised faces and updated by Adam (learning rate 0.01) for 100,000 iterations, with re-projection to the unit sphere after each step. A stochastic mini-batch variant reduces the per-step cost from quadratic in the number of identities to the batch size, and the resulting gradient estimator is provably unbiased. Packing 10K identities takes approximately 6 hours on a single NVIDIA RTX 3090. Image synthesis is delegated to the frozen Arc2Face, and each image is drawn with Gaussian noise around the optimised embedding.

HyperFace generates identities by directly optimising their positions on the 512-d ArcFace unit hypersphere to maximise their pairwise angular distances, which is the classical Tammes packing problem. The same Tammes formulation was used earlier by ID^3 [18] for inter-class identity selection. Released datasets are at 10K, 30K and 50K identities, with the largest containing 3.2M images. Intra-identity diversity has two sources. The first is a Gaussian perturbation added to the packed anchor embedding (scale is 0.01) followed by re-normalisation to the unit sphere, and the second is the diffusion noise input to Arc2Face. While pose, expression or illumination conditioning can be provided to Arc2Face through a ControlNet, intra-class variation still arises implicitly through seeding Arc2Face with a noise around the optimised embedding.

At 10K identities and 50 images per identity (0.5M images total), HyperFace reaches a 5-benchmark average of 89.50%. The 50K-identity 3.2M-image release reaches 91.60%. Image-quality metrics are not reported. One ablation shows that the on-manifold regularisation gallery (StyleGAN, LDM or real BUPT) has only a small effect, while another shows that the HyperFace packing optimisation reaches 1-3 percentage points higher than random hypersphere sampling on most benchmarks.

c. Qualitative and Quantitative Comparison of DisCo and HyperFace to State-of-the-Art Methods

This section contains a qualitative and quantitative comparison of the synthetic data generation methods proposed by Idiap. Table 4-6 contains a qualitative comparison of methods, and Table 4-7 &Table 4-8 contain quantitative comparisons of various synthetic data generation methods, in which datasets are generated at different scales, a face recognition backbone (IResNet) is trained on the generated dataset, and the methods are evaluated on real datasets. Table 4-9 contains evaluation results on RFW for a bias and fairness assessment of DisCo and other state-of-the-art synthetic data generation methods.

Table 4-6 Quallitative comparison of state-of-the-art methods to DisCo and HyperFace.

Method	Source data	Inter-ID strategy	Intra-ID strategy	Input type / control
Arc2Face	GFPGAN-restored WebFace42M (~21M imgs, 1M IDs at 448x448), then FT on FFHQ + CelebA-HQ. Frozen IR-100 ArcFace ID encoder.	PCA on ArcFace embeddings, reject if cos > 0.3 to existing IDs.	Diffusion stochasticity around a fixed pseudo-prompt, pose/expression emerge from the data prior.	512-d ArcFace embedding (zero-padded to 768, substituted for <id> token in CLIP). Optional ControlNet (FLAME normals) for pose/expression.
Vec2Face	1M imgs / 50K IDs from WebFace4M, feature extractor ArcFace-R100 on Glint360K. Identity sampling by PCA on	Sample in MS1MV2 PCA space, accept if cos < 0.3 (1.7% rejected at 300K).	Gaussian noise on anchor with scale in {0.3,0.5,0.7} (40/40/20% mix), keep perturbations with cos > 0.5 to anchor.	512-d feature vector. AttrOP runs gradient descent on the vector against pose (6DRepNet), quality (MagFace) and identity (ArcFace).

© 2024-2027 CERTAIN

	MS1MV2 features.			
Vec2Face+	~4M imgs / 200K IDs from WebFace4M, same feature extractor as Vec2Face.	Direct Gaussian sampling on the 512-dimensional hypersphere, same filter as VecFace.	As Vec2Face, plus a LoRA-based pose adapter conditioned on 5-landmark images. Tuned for intra-class identity consistency.	512-d feature vector + 5-landmark image for the LoRA pose path. AttrOP + LoRA pose.
DisCo	Frozen StyleGAN2 (pretrained on FFHQ), frozen ArcFace IResNet-50 (MS-Celeb-1M), covariate directions via linear SVMs on Multi-PIE.	Langevin applies Brownian repulsion of identities in the face-recognition embedding space, plus pull-back.	Dispersion applies Brownian repulsion in StyleGAN2 W-space, plus a spring attractor in embedding space. DisCo uses the same dynamics, but initialises each intra-class variation differently.	512-d StyleGAN2 W-latent. Implicit covariate control through the 7 Multi-PIE-derived directions (6 L-R poses, 6 illuminations, 5 expressions).
HyperFace	ArcFace 512-d hypersphere (off-the-shelf), frozen Arc2Face for synthesis, on-manifold-regulariser gallery from StyleGAN (default), LDM, or BUPT real.	Tammes packing on the unit hypersphere via Adam, with on-manifold regulariser pulling embeddings toward a gallery.	Gaussian perturbation of the packed anchor, re-normalised to the sphere) plus Arc2Face diffusion noise.	512-d unit-norm ArcFace hypersphere point (passed to Arc2Face). No explicit attribute control.

Table 4-7 Downstream face-recognition accuracy for each method at the same scale of synthetically generated datasets. The last entry allows the comparison of the performance of synthetic datasets with real datasets.

Method	# IDs / # imgs	LFW	CFP-FP	CPLFW	AgeDB	CALFW	Avg
Arc2Face	10K / 0.5M	98.81	91.87	85.16	90.18	92.63	91.73
Vec2Face (HSFace10K)	10K / 0.5M	98.87	88.97	85.47	93.12	93.57	92.00
Vec2Face+ (VFace10K)	10K / 0.5M	99.35	93.56	88.03	94.33	94.17	93.89
DisCo (Langevin+DisCo)	10K / 0.65M	97.07	79.56	76.73	83.38	89.05	85.16
HyperFace	10K / 0.5M	98.50	88.83	84.23	86.53	89.40	89.50
CASIA-WebFace (real ref.)	10K / 0.49M	99.42	96.91	89.78	94.50	93.35	94.79

Draft

© 2024-2027 CERTAIN

Table 4-8 Downstream face-recognition accuracy for each method at different scales of synthetically generated datasets. The last two entries allow the comparison of the performance of synthetic datasets with real datasets.

Method	# IDs / # imgs	LFW	CFP-FP	CPLFW	AgeDB	CALFW	Avg
Arc2Face	20K+40K mix / 1.2M	98.92	94.58	86.45	92.45	93.33	93.14
Vec2Face (HSFace300K)	300K / 15M	99.30	91.54	87.70	94.45	94.58	93.52
Vec2Face+ (VFace100K)	100K / 4M	99.52	95.09	90.23	95.17	94.38	94.88
Vec2Face+ (VFace300K)	300K / 12M	99.57	94.80	90.43	95.17	94.70	94.93
DisCo (Langevin+DisCo)	30K / 1.95M	98.97	83.77	81.52	93.32	93.95	90.31
HyperFace	50K / 3.2M	98.27	92.24	85.60	90.40	91.48	91.60
CASIA-WebFace (real)	10K / 0.49M	99.42	96.91	89.78	94.50	93.35	94.79
WebFace4M (real)	200K / 4M	99.73	98.93	94.93	97.52	96.02	97.23

Table 4-9 Demographic bias due to ethnicity on the RFW benchmark with other State-of-the-Art methods that report a benchmark on RFW.

Dataset (size)	Caucasian	Asian	Indian	African	Avg	Std (pp)
Langevin+DisCo (1.6M)	89.17	82.95	85.43	81.42	84.74	**1.81**
DCFace-1.2M	90.08	82.97	87.05	80.97	85.27	2.66
IDiff-Face (Uniform)	84.92	80.63	81.67	75.65	80.72	2.72
GANDiffFace	76.32	72.85	72.65	66.48	72.03	3.00

4.3.3.2 Intended Use for Compliance Testing

The intended use of the proposed synthetic data generation methods is to generate benchmarking datasets that measure how a face-recognition system's performance varies across the sub-distributions of population and capture condition it will encounter in deployment. The rationale follows the distinction between capture and identity-bound characteristics. In the high-risk, controlled-capture scenarios, the capture (extrinsic) characteristics, namely pose, illumination, expression and resolution, can be constrained largely by the enrolment and acquisition protocol, so the system can be operated within the regime for which it was designed. The identity-bound (intrinsic) characteristics, namely skin tone, ethnicity, gender and age, cannot be controlled, because a deployed system must accept whoever presents themselves. The behaviour of the system across these uncontrollable sub-populations is therefore where compliance testing carries the most weight, and where representative real data is hardest to obtain lawfully. Synthetic generation lets these sub-distributions be produced on demand and in balance.

Compliance testing along these lines has three components, treated in turn below. They are (a) overall *performance*, (b) *demographic fairness* across identity-bound sub-populations, and (c) *bias under capture and soft-biometric factors*. Capture factors are tested even though they are controllable in deployment, both

Draft

because a system's robustness to residual variation within the controlled regime must be characterised, and because that same controllability makes them straightforward to vary synthetically while holding identity fixed.

Each test considers one axis of variation and measures face-recognition verification accuracy along that axis. Axes that test demographic fairness (ethnicity, gender, age) use disjoint identity sets per demographic group, and verification accuracy is computed independently within each group, with the disparity across groups as the metric of interest. Axes that test capture robustness (pose, illumination, expression) use a single identity set in which each identity is rendered at multiple values of the covariate, and verification accuracy is computed for image pairs that share identity but differ on the covariate value.

Table 4-10 Suggested data-generation protocols for face-recognition evaluation. Demographic-fairness axes use identity-disjoint groups, with verification accuracy computed per group and the disparity across groups as the metric. Covariate-robustness axes use a single identity set, with verification accuracy computed across image pairs of the same identity but with different covariate values.

Axis	Type	Levels / groups	Total IDs	Imgs / ID
Ethnicity	Fairness	4 groups (African, Indian, Caucasian, Asian)	4000 (1000 / group)	8–10
Gender	Fairness	2 groups (Male, Female)	2000 (1000 / group)	8–10
Age	Fairness	5 groups (<18, 18–30, 30–50, 50–70, >70)	5000 (1000 / group)	8–10
Pose	Robustness	7 yaw bins (0, 15, 30, 45, 60, 75, 85 degrees)	1000	7
Illumination	Robustness	6 directional conditions	1000	6
Expression	Robustness	6 levels (neutral + 5 Multi-PIE expressions)	1000	6

Such generated datasets can be used for testing face recognition systems for the following aspects.

A. Performance Testing

Performance testing measures the system's overall verification accuracy on a synthetic test set drawn to resemble the deployment population, without stratifying by any single attribute. The reported quantities are the verification accuracy and, at the operating point relevant to the application, the true-positive rate at a fixed false-acceptance rate (TMR@FMR). This establishes the aggregate baseline against which the per-sub-distribution results below are read, since a fairness or robustness gap is meaningful only relative to the overall performance level.

B. Testing for demographic bias and fairness

For demographic-fairness analyses, 1000 identities per demographic group give tight confidence intervals on both the per-group verification metric and the cross-group disparity. With 8 images per identity, each identity contributes $C(8,2) = 28$ same-identity image pairs, so 1000 identities per demographic group yield 28,000 same-identity pair tests per group, plus an equal number of within-group cross-identity pair tests sampled in balance. The 95% confidence-interval half-width on a proportion estimate at $p = 0.95$ from $n = 28{,}000$ pair tests is $1.96\sqrt{p(1-p) \div n} \approx 0.25$ percentage points, and the half-width on the difference between two such proportions is $\sqrt{2}$ times wider, approximately 0.35 percentage points.

A two-proportion z-test on the same $n_1 = n_2 = 28{,}000$ samples lets the disparity between any two demographic groups be tested for statistical significance. The test statistic is $Z = \Delta\sqrt{p(1-p)(1 \div n_1 + 1 \div n_2)}$,

© 2024-2027 CERTAIN

where Δ is the observed cross-group disparity in verification accuracy and p is the pooled accuracy. At $p = 0.95$, the standard error of the disparity is approximately 0.18 percentage points. A disparity of 0.4 percentage points then gives a two-sided p-value of approximately 0.03, and a disparity of 1 percentage point gives $p \approx 6e-8$. The smallest disparity detectable at the $\alpha = 0.05$ significance level is approximately 0.36 percentage points, and the smallest disparity detectable at $\alpha = 0.05$ with 80% statistical power is approximately 0.52 percentage points.

The verification protocol is constructed in two passes. First, for each identity, all $C(n_{img}, 2)$ same-identity image pairs are added to the positive set, which gives 28 positive pairs per identity when $n_{img} = 8$. Second, an equal number of cross-identity pairs is sampled within the same demographic group. For each positive pair, one negative pair is drawn from two distinct identities in the same group. The resulting balanced positive–negative pair set is used to estimate face-recognition verification accuracy per group, and the disparity between any two groups is the difference between their per-group accuracies.

C. Testing for bias under capture and soft-biometric factors

The same statistical argument applies to the capture and soft-biometric axes, with one structural difference. These axes use a single identity set rendered at several values of the covariate, so the quantity of interest is not a cross-group disparity but the change in verification accuracy as the covariate moves away from its reference value. For the pose axis, accuracy is measured on same-identity pairs in which one image is frontal and the other is at each yaw bin, and the bias of the model under pose is the drop in accuracy from the frontal–frontal baseline to each frontal–profile condition. The illumination and expression axes are treated identically, pairing a reference-condition image with each non-reference condition.

With 1000 identities and one reference-to-level pair per identity, each covariate level yields on the order of 1000 same-identity test pairs, giving a 95% confidence-interval half-width on the per-level accuracy of about $1.96\sqrt{p(1-p) \div 1000} \approx 1.4$ percentage points at $p = 0.95$. Detecting a small accuracy drop at a single covariate level therefore requires more pairs per level than the fairness analysis requires per group, so raising the number of images per identity at each level, or the identity count, tightens the interval as $1 \div \sqrt{n}$, where a soft-biometric factor is the axis of interest, such as glasses, facial hair, hair style or makeup, the same single-identity-set, paired-condition protocol applies, with the factor toggled between present and absent while identity is held fixed. Because these factors are at least partly capture-controllable, they can be varied synthetically without the identity drift that changing the identity embedding would introduce.

D. Auditing for Identity Leakage

Generators trained on real face data can produce outputs that match identities in their training corpus. This is a known identity-leakage failure mode. A formal guarantee that a generated dataset is free of leakage is not available for any of the four reviewed methods, but a practical audit can be performed against the generator's training corpus using a pretrained face-recognition backbone.

The audit procedure has three steps as follows.

1- For each generated identity, the ArcFace embedding of each generated image is computed and averaged across the per-identity image set to obtain one anchor embedding per generated identity.
2- The ArcFace embedding of each image in the source dataset used to train the generator is computed.
3- For each generated identity, the maximum cosine similarity between its anchor embedding and any source-image embedding is computed. Generated identities whose maximum source-similarity exceeds a chosen threshold are flagged as potential copies of training identities and dropped from the released synthetic dataset. The Vec2Face+ paper uses a threshold of 0.4 on this similarity.

The source dataset against which to audit depends on the generator. For DisCo, the relevant source is FFHQ, on which the underlying StyleGAN2 was trained. For Arc2Face, the source is the union of the GFPGAN-restored WebFace42M corpus and the FFHQ + CelebA-HQ fine-tuning corpus. For Vec2Face and Vec2Face+, the source is the WebFace4M subset used to train the fMAE generator, with the 0.4 threshold already applied

© 2024-2027 CERTAIN

at dataset release. For HyperFace, the source is Arc2Face's training corpus, since HyperFace itself performs no training.

The audit catches generated identities that match training identities under the face-recognition backbone's similarity metric. It does not catch matches that fall below the embedding threshold but would be recognised as the same person by a human observer, nor matches under a different face-recognition backbone with a different embedding geometry. The audit is therefore a necessary check rather than a sufficient guarantee of synthetic origin.

E. Generating Novel Synthetic Datasets with DisCo and HyperFace for Compliance and Performance Testing

The protocols described in Table 4-10 require identity sets that fall into prescribed demographic groups, or that can be rendered at prescribed covariate levels, in specified quantities. HyperFace and DisCo both produce identities by sampling points in an embedding space, the ArcFace hypersphere for HyperFace and the StyleGAN2 W-space for DisCo, so a target group can be produced either by controlling that sampling directly or by filtering it.

HyperFace synthesises each packed identity with a frozen Arc2Face, so its controllability is that of Arc2Face, and three mechanisms together cover the axes of the protocols. For the capture axes, Arc2Face's ControlNet, conditioned on FLAME face-normal renderings, sets pose and expression directly while the identity embedding is held fixed. For the identity-bound and soft-biometric axes, where no direct control exists, an attribute classifier is first trained on real labelled faces for the target property, for example skin tone, gender, age or eyewear, taking an ArcFace embedding or the rendered image as input. Each candidate identity is then generated, scored by the classifier, and retained only if its predicted attribute matches the target group. This rejection sampling is repeated until every group in Table 4-10 reaches its identity quota.

DisCo follows the same approach without the ControlNet step. Pose, illumination and expression are set directly by the seven Multi-PIE covariate directions in W-space. For the identity-bound and soft-biometric axes, an attribute classifier is fit on real data in W-space or in the ArcFace embedding space, which is the same kind of linear fit DisCo already uses to obtain its covariate directions, and the latents are reject-sampled until each target group is filled. Because the classifier acts on the latent before any image is rendered, the filtering for DisCo is cheap, and only the retained identities are passed through the generator.

The procedure fixes the marginal distribution of each attribute by construction, at the cost of generating and discarding candidates in proportion to how rare the target group is under the generator's prior.

The instructions to run HyperFace or DisCo themselves are given in Section 4.3.2.2.

4.4 ENERGY

4.4.1 Overview: A Unified Synthetic-Data Framework for Energy

This section presents the CERTAIN energy-domain contribution as a single synthetic-data capability developed along a common axis. The shared problem is that the energy data most useful for training and evaluating AI systems includes granular smart-meter time-series, which is privacy-constrained and, for time-series of this kind, cannot be safely opened through pseudo-anonymisation alone, which blocks open compliance evaluation. Two complementary synthesis methods address this. The first, source-anchored re-synthesis, starts from a real Austrian energy-community ecosystem (EMPOWER), irreversibly destroys all personally identifiable information at the synthesis boundary, and re-synthesises a GDPR-compliant, openly publishable dataset that preserves the specific regulatory and operational mechanics of Austrian energy communities. The second, generative synthesis from public priors, uses a set of learned, diffusion-based generators trained on publicly available datasets that require no sensitive source and can extrapolate beyond it including new countries, weather regimes, household compositions, and missing or unseen attributes.

© 2024-2027 CERTAIN

The two regimes occupy opposite ends of a single spectrum, ranging from high-fidelity reproduction of one real regulated ecosystem to flexible generative extrapolation beyond any single source, and are deliberately complementary rather than redundant. Together they cover what a compliance tester actually needs: realistic in-distribution data that faithfully reflects a production environment, and controllable out-of-distribution, cross-country, and rare-scenario data for stress-testing. Table 4-11 below positions the two regimes against the dimensions that matter for choosing between them.

Table 4-11 Positioning of the two synthesis regimes

Dimension	Source-anchored re-synthesis	Generative synthesis from public priors
Source data	One real Austrian EEG ecosystem (EMPOWER); PII destroyed at the synthesis boundary.	Publicly available datasets only (e.g. Pecan Street, OPSD, UK-DALE); no sensitive source.
Extrapolation	Within the source ecosystem's regime; bounded by single-source coverage.	Beyond the source: adaptation to new countries, weather regimes, household types, and missing features.
Fidelity basis	Statistical match to the real ecosystem, validated per cluster.	Physical and statistical plausibility learned from public data, enforced by dedicated loss terms.
Privacy mechanism	Full re-synthesis with irreversible PII destruction; meets the GDPR Recital 26 standard by construction.	No personal source data; uses only publicly available datasets and generation methods.
Regulatory anchoring	EAG/EEG-compliant entities, OBIS-coded measurements, Austrian identifier formats.	Country-agnostic; conditioned on configurable priors and household metadata.
Primary role for testing	In-distribution benchmark for Austrian EEG operations.	Out-of-distribution, cross-country, and rare-scenario generation.

To keep the two regimes legible as one framework, this section adopts a shared vocabulary where the approaches meet. The terms "metering point" and "household" both denote the individual generation or consumption unit; "cluster" and "archetype" both denote a homogeneous group sharing one synthesis model; and "physical-plausibility constraint" denotes the enforcement (by post-hoc clipping or by a penalty loss) that generation and consumption remain non-negative and within plausible capacity. Community- and grid-level quantities use the notation introduced in the methodology section ($G_{t,ec}$, $C_{t,ec}$, $a_{t,ec}$, $u_{t,ec}$).

4.4.2 Domain description

Both contributions describe the same domain including multi-scale energy generation and consumption whose most useful data is privacy-constrained building up from two complementary vantage points: the regulated Austrian energy-community setting re-synthesised from a real ecosystem, and the broader multi-scale energy environment modelled generatively from public data.

The source-anchored perspective:

The energy domain centres on the Austrian energy community (EEG) framework established under the Erneuerbaren-Ausbau-Gesetz (EAG, BGBl 150/2021), a law defining this framework. Austria has committed

© 2024-2027 CERTAIN

to 100 % renewable electricity by 2030 and full climate neutrality by 2040, and the EEG construct is one of the central legal instruments of this transition. By 2026, more than 2,000 such communities are operational, in which households, small businesses, and producers jointly generate and consume renewable electricity through shared photovoltaic, hydropower, and (less commonly) wind installations. Within each community, smart meters generate continuous 15-minute interval readings per metering point that are weighted by individual participation factors, aggregated into community-level coverage and surplus indicators, and reconciled by the network operator into officially billable values.

This data layer is simultaneously the most valuable and the most sensitive substrate for AI applications in the Austrian energy transition. It is highly granular, persistent, geographically anchored, attributable to identified natural persons, and behaviourally informative. Characteristic appliance signatures, household routines, vacation patterns, and shift schedules are all recoverable from 15-minute consumption traces. Pseudo-anonymisation by stripping names and identifiers is documented in the literature to be insufficient for time-series of this kind, and the GDPR Recital 26 standard for anonymity is correspondingly difficult to meet. As a consequence, the very data that would best support open research in the EEG domain cannot, in practice, be openly released by EEG operators, even when they wish to do so. As the academic partner within the EMPOWER energy pilot, this contribution addresses the gap by producing a fully re-synthesised, GDPR-compliant, openly publishable EEG dataset.

The synthetic-data objective in this domain is therefore not only to provide more training data, but, most importantly, to provide publishable training and evaluation data that preserves the structural and statistical properties of real EEG operations while severing all correspondence to natural persons. This requires that the synthesis preserve four properties simultaneously, none of which is addressed by existing public datasets in combination: (i) community-level mechanics, including surplus distribution, coverage allocation, and participation factor schedules; (ii) Austrian regulatory alignment, with EAG-compliant entities and OBIS-coded measurement types; (iii) full re-synthesis with no PII, rather than pseudo-anonymisation of source records; and (iv) open licensing for unconstrained downstream research and tooling.

The primary AI applications enabled by such a dataset span two families. First, demand and generation forecasting, both at the individual metering-point level and at the community-aggregate level (community generation $G_{t, ec}$, community consumption $C_{t, ec}$, self-coverage rate $a_{t, ec}$, surplus rate $u_{t, ec}$), with weather as an exogenous input. Second, distribution-algorithm and participation-factor analysis — evaluating how alternative participation-factor schedules, community memberships, or distribution rules would have affected coverage and surplus, supporting both research on EEG fairness and operational decision support for community managers. In addition, the dataset supports auxiliary tasks such as clustering and archetype discovery, anomaly detection, and the evaluation of energy-policy scenarios under controllable conditions.

Where the source-anchored view stays within one regulated national setting, the generative view generalises the same domain across scales and countries.

The generative perspective:

Using generative methods, the energy domain shifts from localized settings toward the modelling of universal behavioural and physical dynamics governing multi-scale energy consumption and generation. Crucially, this perspective is not constrained to a single national context but can be adapted to different geographical, demographic and climatic conditions. The energy domain focuses on modelling complex interactions across both the micro (individual household) and macro (grid and national) levels. Energy systems are inherently multi-scale, highly dynamic, and influenced by a wide range of external factors, including weather conditions, user behaviour, market signals, and the increasing penetration of renewable energy sources. Accurately capturing these interactions is essential for developing reliable and scalable energy intelligence solutions.

The objective of synthetic data generation in this context extends beyond high-fidelity replication of a single national dataset towards multinational generation and scenario diversification. It addresses the limitations of real-world datasets, which are often constrained by privacy concerns, limited accessibility, fragmentation, and insufficient temporal or spatial coverage. This is achieved by leveraging a multinational training pool to overcome these limitations, and enable the generation of diverse, physics-plausible, and geography-agnostic energy datasets. Rather than merely reproducing observed data, the framework aims to capture the underlying behavioural and physical dynamics governing energy consumption and generation across different contexts.

Draft

© 2024-2027 CERTAIN

In particular, household-level data is typically scarce and sensitive, while large-scale grid data lacks the granularity needed to model individual behaviours. Synthetic data generation enables the creation of complete, consistent, and physically plausible representations of realistic energy environments, bridging this gap by preserving both fine-grained consumption patterns and large-scale system dynamics.

The primary AI applications enabled by this generative paradigm focus on cross-regional robustness and scalability. First, **multi-scale, cross-border demand and generation forecasting**, enabling the development of predictive models that remain resilient when deployed across shifting geographical and climatic baselines. Second, **macro-level grid stress-testing and long-term infrastructure planning**, which utilizes the generated data to simulate how large-scale, interconnected grid systems respond to unprecedented renewable penetration or extreme, multi-region weather events. Finally, it powers **geography-agnostic residential electricity load profiling**, allowing researchers and operators to simulate localized demand-side management, optimize distributed energy resources (DERs), and benchmark transnational energy policies under highly controllable, simulated conditions.

Forecasting as a shared, multi-modal capability: Across both contributions, forecasting is treated as a single shared capability rather than two separate tasks. Instead of a forecasting-only pipeline, the combined framework defines a broader, multi-modal forecasting and data-completion capability. The source-anchored dataset provides a realistic in-distribution Austrian benchmark for demand and generation forecasting at the metering-point and community-aggregate levels ($G_{t,ec}$, $C_{t,ec}$, $a_{t,ec}$, $u_{t,ec}$), while the generative method extends this capability to predict missing features for new entries within communities, to support the creation of synthetic community profiles, and to generate communities for other countries with different characteristics and weather conditions. Framed this way, forecasting, imputation, and conditional generation become one capability spanning in-distribution evaluation and out-of-distribution extension, rather than overlapping descriptions of the same task.

4.4.3 Source Data Characteristics

The two contributions draw on deliberately different source data, which is the root of their complementarity: the source-anchored method models one real, regulated Austrian ecosystem, while the generative method draws on a broad set of public datasets. The two are described in turn.

Source-anchored re-synthesis: the EMPOWER ecosystem:

The source-anchored synthesis pipeline is grounded in a single source ecosystem: the EMPOWER ecosystem operated by the pilot's academic partner. This ecosystem comprises approximately 30 real Austrian energy communities active during the 2025 baseline year, each contributing 15-minute interval readings for all of its metering points; two communities additionally provide partial 2024 data. The total source-data volume is on the order of millions of metering-point-quarter-hours per year. Using a single coherent source ecosystem is an explicit scope decision: it provides a unified regulatory and operational frame (a single legal regime, a single network-operator distribution algorithm, a single set of OBIS measurement conventions) at the cost of single-source bias, the implications of which are discussed in §4.4.5 and in the project limitations register. No raw EMPOWER data is exposed beyond the source partner; only PII-stripped time-series and aggregated structural summaries cross the synthesis boundary.

The dataset type is a relational time-series ecosystem rather than a single table. It combines three categories of information: (i) static descriptive data of energy communities, participants, and metering points, including their structural relationships; (ii) time-bounded relational data linking metering points to communities via participation factor schedules; and (iii) a high-volume time-series fact table of 15-minute interval energy measurements, with each row tagged by a measurement type from the standardized OBIS code system. Around 30–40 distinct fields are present across the schema, of which the time-series fact table alone contributes the bulk of the row volume.

The schema entities and their core fields are as follows. The `energy_community` entity holds static EEG attributes (surrogate and public identifiers, name, geographic coordinates of a community reference point). The `participant` entity describes EEG members (surrogate and public identifiers, member type — private or business, contact metadata). The `metering_point` entity describes the smart-meter unit (a structurally valid Austrian metering-point number, a direction enumerator $\{C, G\}$ for consumer versus generator, and for

© 2024-2027 CERTAIN

generators an energy-source enumerator (Solar, Water, Wind). The `object` entity is a common parent for participants and metering points carrying address attributes (street, number, postal code, city) and geographic coordinates. The `participation_factor` entity is a time-bounded relation linking a metering point to an energy community with a percentage value modelled as integer in $[0, 100]$ and a `valid_from` timestamp marking the start of the schedule window. The `energy_data` entity is the fact table: a row references a metering-point `object_id`, an `organization_id` (the EEG context), a `period_begin` timestamp persisted in UTC, a `period_duration` of 15 minutes, a non-negative `energy_value` in kWh, and a `measurement_type`. Reference data (weather, market prices) are part of the conceptual model for forward compatibility but are outside the synthesis scope of the initial release.

Feature types span all three of the standard families. Continuous features include the energy values themselves and the geographic coordinates. Categorical features include participant type, energy direction, and energy source. Ordinal/integer features include the participation factor (percentage value as integer) and the surrogate identifiers. Identifier features (metering-point number, participant number, community number) are structurally constrained strings that conform to the Austrian formats while being entirely synthetic at publication time. Temporal features are confined to the `energy_data` and `participation_factor` tables, where they take the form of timezone-aware UTC timestamps with explicit period durations.

The `measurement_type` enumeration is itself a domain-specific feature derived directly from the Austrian regulatory schema. Five families are represented in the initial release, each with its OBIS code: Weighted Measured Consumption (WMC, OBIS 1-1:1.9.0 G.01T), Weighted Measured Generation (WMG, OBIS 1-1:2.9.0 G.01T), Weighted Surplus Generation (WSG, OBIS 1-1:2.9.0 P.01T), Community Coverage (CC, OBIS 1-1:2.9.0 G.03), and Community Potential (CP, OBIS 1-1:2.9.0 G.02). The first two are direct measurements weighted by participation factor; the latter three are derived community-level quantities computed by the network-operator's distribution algorithm.

Temporal aspects are central to the source data. All time-series are sampled at a fixed 15-minute resolution, which is the legally prescribed resolution for energy communities in Austria and is the resolution preserved end-to-end through synthesis. The default temporal scope of the source data and of the synthesised release is calendar year 2025 (35,040 quarter-hour intervals per metering point), with limited 2024 coverage available for two communities to support inter-annual extension experiments. Participation factor schedules introduce a second, slower temporal dimension: under the current legal regime they update on a daily cadence, with planned legislation moving to an hourly cadence; the schema supports both. Daylight Saving Time transitions are handled natively by persisting timestamps in UTC and resolving display-time conversions at query time.

In contrast to the single regulated ecosystem above, the generative method learns from a broad pool of public datasets, which is what enables its later extrapolation across countries and conditions.

Generative synthesis: public-dataset characteristics:

To train the generative models, the framework extracts diverse characteristics from several well-known real-world public datasets, including Pecan Street, London Smart Meters, UK-DALE, Open Power System Data (OPSD), CoSSMic, UCI Household, and the WiDS Datathon 2022 dataset. To achieve cross-border generative capabilities, the models are trained on geographically diverse energy and weather data spanning multiple countries, climates, and household types. Rather than learning country-specific behaviours, the framework aims to capture the underlying physical and behavioural dynamics governing energy consumption and generation. This enables the model to generalize across different geographical regions through configurable weather, demographic, and household metadata.

The source data consists of a combination of high-frequency multivariate time-series measurements, such as smart meter readings and grid-level data and includes tabular metadata describing household and building characteristics. The datasets provide measurements at multiple temporal resolutions, ranging from sub-minute frequencies (e.g., 1 Hz and 16 kHz) and one-minute intervals to 15-minute, hourly, daily, and annual aggregations, often spanning several years. This combination enables the model to capture complex interdependencies between energy consumption, weather conditions, household characteristics, and temporal dynamics across multiple scales. Consequently, the sampling frequency can be adapted to the requirements of different use cases while allowing the model to learn and preserve both short-term consumption patterns and long-term seasonal trends through multi-resolution temporal learning. Another important aspect is the

© 2024-2027 CERTAIN

richness and diversity of the available features and their types. The datasets include a wide range of continuous, categorical, and derived features. Core features include global and local timestamps, household identifiers, country or regional location, building energy class, and energy consumption and generation metrics. In addition, meteorological features are extensively used, such as temperature, humidity, wind speed, precipitation, and cloud cover. During preprocessing, additional derived features are created to enrich the training process. These include Heating and Cooling Degree Days (HDD/CDD), which capture thermal comfort demand, as well as sinusoidal encodings (using sine and cosine functions) to represent periodic time variables (e.g., hour of day, day of week) in a continuous form. By combining raw measurements with engineered features, the model is able to learn more meaningful relationships and capture both physical and behavioural patterns within the energy domain.

Before moving to the next section, it is important to provide additional details regarding the main datasets used in this work:

- The **Pecan Street dataset** provides total energy consumption aggregated from all electrical sources within each individual home, covering 417 households in the USA. The data spans a full calendar year, from 1 January 2017 to 31 December 2017, with an hourly resolution. This dataset enables even relatively simple models to capture daily, monthly, and yearly trends, and to implicitly extract patterns related to weather conditions, even though weather features are not directly provided as inputs to the model.

- The **Smart Meter dataset in London** and the **UK-DALE dataset** both provide information on household energy consumption in the UK. The former includes data from approximately 5,500 households, with a sampling frequency of 30 minutes to 1 hour, spanning from November 2011 to February 2014. Similar to the Pecan Street dataset, its multi-year duration allows simple models to capture seasonal trends and indirectly learn weather-related patterns, even without explicit weather inputs. The latter dataset captures both individual appliance-level power demand and aggregate whole-house mains consumption for 5 households. It offers very high sampling frequencies (1 Hz and up to 16 kHz for mains, and 6-second intervals for appliances) over a multi-year period. This makes it particularly useful for synthetic energy data generation, especially when developing complex ML models that aim to accurately represent how individual smart home devices (e.g., heating, cooling, and kitchen appliances) contribute to the overall household energy profile.

- The **Open Power System Data (OPSD)** platform provides macro-level energy data from various European countries, primarily focusing on national and regional grid operations. This includes aggregated electricity consumption, wind and solar generation, and installed capacities, and associated meteorological variables such as temperature, wind speed, solar irradiation, and other weather observations, spanning multiple years with high-resolution measurements (typically 15-minute to hourly intervals). OPSD is considered highly curated and standardized, making it valuable for synthetic energy data generation, particularly for training ML models aimed at predicting macro-level energy surpluses, simulating long-term grid planning scenarios, or learning large-scale renewable generation patterns. In addition, OPSD includes a limited dataset with household-level consumption data, although this is not its primary focus.

- Finally, the **UCI Household Power Consumption dataset** provides detailed measurements for a single residential household in France. It includes total household demand as well as the power consumption of three specific sub-zones. The dataset has a 1-minute resolution and spans nearly four years, from December 2006 to November 2010.

4.4.4 Synthetic Data Generation Methodology

The methodology is presented as one pipeline whose stages are defined by function, each realised by two complementary methods described side by side: a source-anchored re-synthesis of a real regulated ecosystem, and a learned generative architecture. This enables a cross-learning set of specialised generators

© 2024-2027 CERTAIN

rather than a single monolithic model trained on publicly available data. The pipeline is modular and methodology-agnostic at its core: the dataset is defined by its schema, scope, and quality-acceptance gates rather than by any single algorithm, and a simple-to-complex selection strategy is followed within each stage, promoting machine-learning methods only where simpler statistical baselines fail to meet the fidelity, utility, and privacy criteria.

Source preparation. The pipeline begins by securing a privacy-safe input. In the source-anchored method, the raw ecosystem data, which carries full personal data, is reduced into two artefacts that cross the synthesis boundary: a PII-stripped time-series keyed only by anonymised metering-point identifiers, and a structural skeleton of counts, types, and aggregate statistics (generation/consumption ratios per community, regional distributions, participation-factor histograms) that informs design but contains no individual-level information. The real-to-anonymised mapping is held only by the source pilot and never exposed, so PII destruction is irreversible at the boundary and all later stages are by construction free of source PII. The generative method requires no equivalent step, since it trains only on publicly available datasets and processes no personal data. However, several preprocessing and feature-engineering steps are applied to enrich the training data and improve model performance. As described previously, these transformations include cyclical sine/cosine encodings of temporal variables, boolean indicators for weekends and public holidays, weather-derived variables, and additional engineered features such as Heating and Cooling Degree Days (HDD/CDD). These enrichments enable the model to better capture temporal, behavioural, and weather-dependent energy consumption and generation patterns.

Temporal backbone. Both methods rest on a reproducible, timezone-aware time axis at a fixed resolution. The generative method builds this explicitly with a deterministic, non-learned temporal generator: given a start date, end date, sampling frequency, and timezone, it constructs a sequence of timezone-aware timestamps (using pandas date ranges) that resolves Daylight Saving Time natively, and extracts calendar-derived features (floating-point hour, day-of-week, month, and day-of-year, plus boolean weekend and national-holiday indicators) together with sinusoidal (sine/cosine) encodings of the hour, day-of-week, and day-of-year so that downstream diffusion models can interpret cyclical structure. The source-anchored method inherits the same backbone from the regulated data: all series are sampled at the legally prescribed 15-minute resolution and persisted as UTC timestamps with explicit period durations, which likewise resolves Daylight Saving Time by construction and is preserved end-to-end through synthesis.

Conditioning and metadata. Each unit and community is described by static and contextual attributes that condition synthesis. In the source-anchored method these come from the structural skeleton and from synthesised master data: structurally valid Austrian metering-point numbers, participant names, addresses, and email addresses generated from faker-style libraries with Austrian locale data and open address registers (e.g. Statistik Austria), each verified so that no synthetic identifier or address coincides with a real one. Time-bounded participation-factor schedules link metering points to communities and are synthesised under hard legal constraints (integer values in [0,100], a per-metering-point sum no greater than 100 across communities, realistic daily or hourly update cadences, and stable temporal regimes rather than random jitter) using random sampling within constraints, sampling from the empirical schedule distribution, or an optimisation procedure. Geographic linkage is treated as a mandatory conditioning feature so that ambient weather can be retrieved for synthetic locations: each synthetic community inherits its source region without inheriting any real address, and an experiment quantifies how weather coupling weakens as the synthesis pool is widened across one, two, three, or more clustered source communities. In the generative method, conditioning is produced by two dedicated generators. A household-metadata generator samples static building attributes from a directed acyclic graph of conditional categorical and Bernoulli distributions plus a continuous Gaussian latent variable $z \sim N(0,I)$, configurable through priors over countries, regions, Energy Performance Certificate (EPC) classes, dwelling types, floor areas, and technology-adoption rates (photovoltaics, electric vehicles, heat pumps); for each household it samples top-down (country, then region, then EPC class and dwelling type, then technology presence via Bernoulli draws conditioned on EPC class), then a latent vector z and a multinomial behavioural-archetype label. A weather generator supplies meteorological context as a multivariate conditional diffusion model that jointly synthesises temperature, humidity, wind speed, and cloud cover; conditioned on the temporal tokens and categorical region indices, a dedicated conditioning encoder adds their combined projection directly to the input representation before it enters the main backbone. Operating under a cosine denoising schedule, this backbone processes the pre-conditioned input through a sequence of residual blocks that cleanly chain dilated 1-D causal convolutions for short-range dependencies, Fourier blocks (FFT, learnable complex weights, inverse FFT) for recurring patterns, and cross-variable attention across the weather variables to preserve their physical correlations.

© 2024-2027 CERTAIN

Per-unit synthesis (per cluster and per household): At the core, each method generates the individual generation and consumption series. The source-anchored method first clusters metering points into homogeneous groups within which one model is fitted: generators are separated into PV and hydro (which differ sharply — PV is strongly diurnal and summer-peaked, hydro is smoother across day and year with non-zero night generation), and consumers into residential, commercial, and industrial along weekend/weekday ratio, day/night ratio, base-load fraction, and annual consumption, using either interpretable rule-based thresholds or k-means / Gaussian-mixture models, validated against the participant-type field, by t-SNE or UMAP separation of prototypes, and by cross-validation stability. A synthesis model is then fitted per cluster from a spectrum of candidates (statistical sampling of per-cluster marginals); a profile-plus-residual block bootstrap that separates a deterministic per-cluster time-of-day profile from a bootstrapped stochastic residual preserving short-range autocorrelation; Gaussian or vine copulas for joint dependence; variational autoencoders; conditional GANs (e.g. TimeGAN, RCGAN); and denoising-diffusion models judged against seven criteria: distributional fidelity (Kolmogorov–Smirnov / Wasserstein), autocorrelation preserved to at least 24-hour lag, cross-series correlation within a community, training-data efficiency at the per-cluster scale, interpretability and auditability, computational cost, and suitability for high row counts. The generative method instead synthesises each household and the grid with coupled conditional denoising diffusion probabilistic models (DDPMs). A Household Context Composer fuses the temporal tokens, weather sequences, and household-metadata vectors into a dense context, while a hierarchical multi-scale denoiser decomposes generation into annual, weekly, daily, and fine scales using Residual Temporal Blocks that combine FiLM and cross-attention conditioning, temporal convolutions with self-attention for local dynamics, and FFT-based spectral processing for periodicities. A Grid Context Composer processes temporal tokens, geographically aggregated weather, country embeddings, and a bottom-up estimate of household demand, driving a multi-resolution encoder–decoder with skip connections. That emphasises Fourier blocks to preserve low-frequency seasonal structure and a dual-output head producing total consumption and generation; a folding technique reshapes consecutive time steps into the channel dimension for efficiency on long sequences.

Cross-scale and community-level assembly: Both methods reconcile the individual series into community- or grid-level quantities, one by a deterministic regulatory formula and the other by a learned cross-scale correction. In the source-anchored method, the synthesised metering points are assembled into archetype EEGs (the choice between a single representative **Demo-EEG**, three to five archetypes anchored on the partner-community profiles, or a parametric archetype generator, and the Austrian network-operator distribution algorithm is applied deterministically to compute the community-level weighted, coverage, and surplus quantities. Given sets of generation metering points E, consumption metering points V, communities EC, and time points T, with raw generation g_{e, t}, raw consumption c_{v, t}, and participation factors pf_{z, t, ec}, the algorithm produces:

$$\mathrm{g}_{e,t,ec} := \mathrm{g}_{et} \cdot \frac{pf_{e,t,ec}}{100}$$

$$\mathrm{c}_{v,t,ec} := \mathrm{c}_{v,t} \cdot \frac{pf_{v,t,ec}}{100}$$

$$G_{t,ec} := \sum_{e \in E} g_{e,t,ec}$$

$$C_{t,ec} := \sum_{v \in V} c_{v,t,ec}$$

$$a_{t,ec} := min\left\{1 \middle| \frac{G_{t,ec}}{C_{t,ec}}\right\}$$

$$u_{t,ec} := max\left\{0 \middle| \frac{G_{t,ec} - C_{t,ec}}{G_{t,ec}}\right\}$$

© 2024-2027 CERTAIN

These derived quantities populate the corresponding `measurement_type` rows in the `energy_data` table. Stage 4 is a deterministic computation given the synthesised raw g and c series and the participation-factor schedules; no further generative methodology is required at this stage.

In the generative method, the cross-scale step is handled by a learned coupling mechanism rather than a fixed formula: individual synthetic household loads are aggregated into a bottom-up estimate of regional demand, which is passed to the grid model as a conditioning signal. Instead of enforcing strict equality, the grid model generates grid-level consumption and generation independently; a consistency regularisation loss softly penalises large deviations between the grid model's consumption output and the bottom-up estimate, encouraging statistical alignment across scales without constraining the grid model to exactly reproduce the household sum. The two assembly mechanisms are thus analogous and both reconcile individual series into community- or grid-level quantities while one applies a deterministic regulatory algorithm and the other a soft, learned cross-scale correction.

Constraint enforcement: Both methods guarantee physical and structural validity, by post-processing in one case and by training objectives in the other. The source-anchored method clips outputs to hard physical bounds (no negative generation or consumption, and no super-peak values beyond plausible per-meter capacity) and enforces schema-level constraints (non-negative energy values, UTC-typed timestamps, OBIS-conformant measurement types, and structurally valid Austrian identifiers), cross-table referential integrity, and internal and external identifier uniqueness; the whole pipeline is captured in a single configuration under fixed random seeds with DVC-tracked lineage, so that the synthesis is end-to-end reproducible by third parties from the same inputs. The generative method, by contrast, addresses data quality upstream during a dedicated data preprocessing stage: rather than removing dropping rows, a quality-assurance pipeline executes rule-based checks for impossible states such as negative consumption or out-of-range humidity, and Interquartile-Range (IQR) and Median-Absolute-Deviation (MAD), and rolling Z-score filters to flag and mark extreme anomalies. Downstream physical and structural validity are subsequently enforced during model training through a set of specialized losses aggregated with configurable weights. The foundational diffusion loss is the mean squared error between the true and predicted noise across all DDPMs; a spectral loss compares FFT magnitudes (L1 or L2, log-scaled) to preserve dominant periodicities; a statistical loss matches batch means, variances, and quantiles to maintain realistic variability and tails; a physical-plausibility loss applies a one-sided ReLU penalty to denormalised values that breach predefined lower or upper bounds; a cross-correlation loss matches per-batch Pearson correlation matrices (mean squared Frobenius norm) to preserve inter-variable weather dependencies; a conditional-consistency loss matches weather-response descriptors extracted by a frozen pretrained regressor so that generated profiles react correctly to their context; and a coupling loss penalises the residual between aggregated household output and the grid estimate. Reproducibility in the generative method is likewise controlled through configurable priors and fixed random seeds.

Draft

© 2024-2027 CERTAIN

Generative Synthesis Architecture

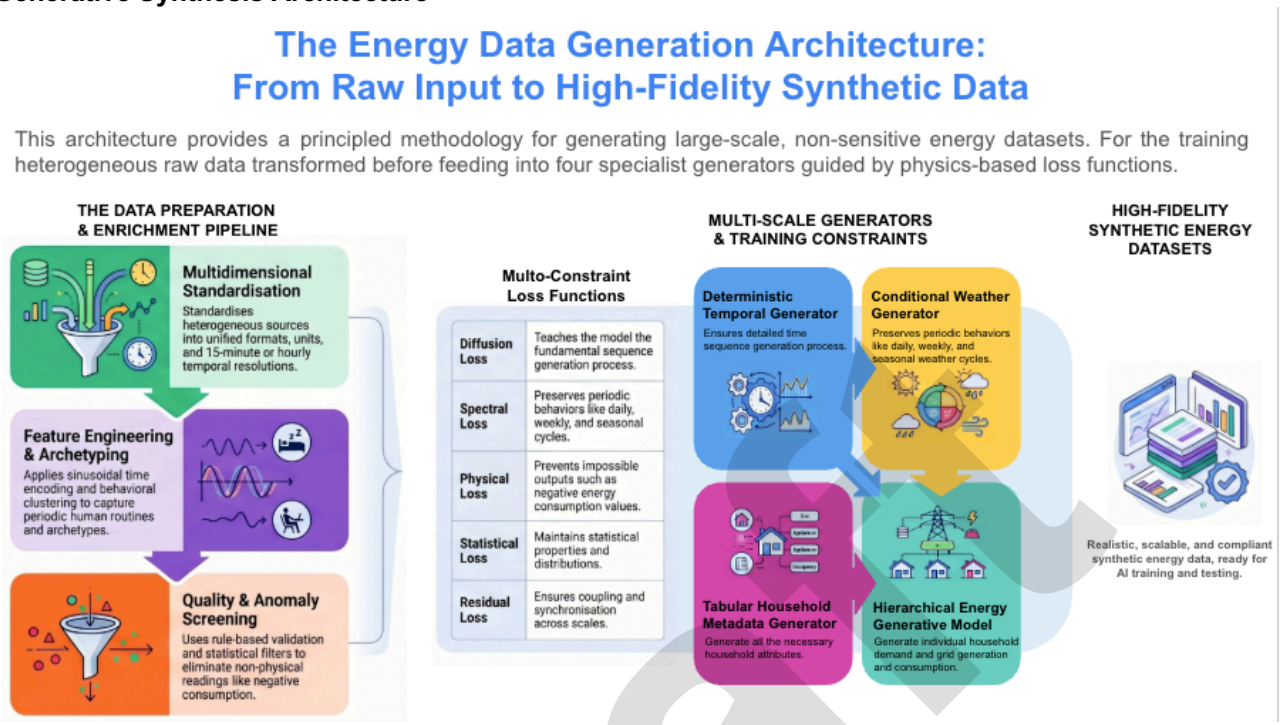

Figure 4-12 Energy Data Generation Architecture

4.4.5 Intended Use for Compliance Testing

Both contributions target the same compliance goal by producing data on which AI systems, and forecasters in particular, can be evaluated before deployment; and they divide the work along the spectrum established above. The source-anchored dataset supplies a rigorous in-distribution evaluation programme, while the generative models extend testing to conditions no single real source contains.

In-distribution evaluation on the source-anchored dataset:

The dataset is designed as a substrate for evaluating AI systems that operate over Austrian energy-community data: forecasting, visualization, and optimization simulations for optimal participation factors derived from these forecasts. The dataset's compliance-testing role rests on a dual claim: on the one hand, it is realistic enough to act as a valid stand-in for production EEG data in evaluation; on the other hand, it is fully re-synthesised and therefore safe to use openly in evaluation reports, demonstrators, and published artefacts. Whether each of these claims holds in practice is itself part of the evaluation programme, which is structured along three dimensions (fidelity, utility, and privacy) each with its own metrics, reporting cadence, and acceptance gates.

Fidelity testing: Fidelity is reported per cluster (PV, Hydro, residential, commercial, industrial) and per derived community-level quantity (G, C, a, u, CC, CP). Distributional fidelity is captured via the Kolmogorov–Smirnov statistic, the Wasserstein distance, and first- and second-moment deltas. Temporal fidelity is captured via autocorrelation-function comparison up to 24-hour lag, daily-profile RMSE (mean by hour-of-day), and weekly-profile RMSE (mean by weekday). Cross-series structure is captured via the correlation-matrix delta over the metering points within an EEG. Aggregate fidelity is captured by community-level deltas on the published distribution-algorithm outputs. Because the dataset's row count runs into the millions, p-values from significance tests are uninformative (essentially zero regardless of practical similarity); reporting therefore emphasises effect-size statistics directly.

© 2024-2027 CERTAIN

Utility testing — the AI-systems compliance bar: Two task families operationalise utility:

1. Forecasting benchmark: A standard load-forecasting model (LSTM or temporal-fusion-transformer baseline) is trained on synthetic data and evaluated on synthetic and real holdouts. The forecasting-error ratio between the synthetic-trained and real-trained variants is the headline number. This tests whether models trained on the synthetic dataset generalise to real-world EEG operations.

2. Distribution-algorithm replay: The Austrian network-operator distribution algorithm formalised in §4.4.4 is run on synthetic raw inputs, and the resulting community-level quantities are compared in distribution to those produced by running the same algorithm on the source raw inputs. Indistinguishability under defined effect-size thresholds confirms that the dataset preserves community-level operational behaviour.

Robustness and stress testing: Beyond aggregate fidelity and utility, the dataset is designed to support robustness evaluation along several axes:

- **Temporal robustness:** the dataset's calendar coverage spans a full year of 35,040 quarter-hour intervals, allowing AI systems to be evaluated across daily, weekly, and seasonal regimes within a single substrate; two source EEGs additionally provide partial 2024 data for inter-annual stability tests.
- **Compositional robustness:** the EEG archetypes (single representative community, multiple fixed archetypes anchored on partner profiles, or a parametric archetype catalogue) allow AI systems to be evaluated across a range of community compositions and sizes.
- **Operational robustness:** the participation-factor schedule synthesis supports both the current daily cadence and the planned hourly cadence, allowing systems to be evaluated under both legal regimes. Stress testing with rare scenarios (extreme weather days, plant outages, abrupt schedule changes) is supported in principle by direct injection into the deterministic Stage-4 assembly; a curated rare-scenario catalogue is identified as a future-extension item.

Bias and coverage testing: The dataset's clustering structure (residential, commercial, industrial; PV, Hydro) and per-cluster reporting expose heterogeneity that single-score datasets obscure. AI-system bias can be probed by comparing performance across clusters, e.g. *"Does a forecasting model trained at the dataset level systematically under-perform on industrial profiles?"* The schema's archetype and EMPOWER EEG-anchors further support coverage-style bias evaluation across community profiles that real Austrian EEGs span. A known limitation, explicitly disclosed, is that the dataset derives from a single source ecosystem (EMPOWER); claims about Austrian EEGs in general are therefore bounded by the regional, demographic, and operational coverage of that ecosystem.

Privacy testing: Privacy metrics confirm that the dataset's "fully re-synthesised, no PII" claim holds operationally: a record-level distance check confirms that for every synthetic time-series, the distance to its nearest neighbour in the source dataset exceeds a threshold under a defined metric (rejecting accidental memorisation by the synthesis model); a membership-inference proxy test verifies that a classifier cannot perform better than chance at distinguishing source-set from non-source-set metering points; and identifier-integrity checks verify that no synthetic master-data value (address, name, metering-point number, email) coincides with a real one. These checks are part of the synthesis pipeline's gating logic and must pass before publication.

Validity for the principal compliance use case: A central compliance question is whether a forecasting model trained, evaluated, and demonstrated on the synthetic dataset can be deployed against real EEG data without surprise. The answer is positive to the extent that the three quality dimensions are satisfied simultaneously, and the per-cluster reporting, the dual-substrate query benchmark, and the distribution-algorithm replay are designed jointly to provide that evidence. The acceptance thresholds against which each metric is judged depends on three negotiated parameters: the chosen synthesis methodology (different methods have different inherent ceilings on certain metrics), the priority weighting of fidelity vs. utility vs. privacy, and the agreed comparator (acceptance against the source EEGs, against published synthetic-data

Draft

baselines, or against an absolute domain-derived target). The companion data paper published in a future step will report the chosen thresholds, the achieved values, and the acceptance / rejection decisions transparently.

Complementing this in-distribution programme, the generative method stresses AI systems against conditions that lie outside any single real source.

Out-of-distribution and stress evaluation with generative data:

The generated synthetic datasets are designed to rigorously evaluate AI systems, particularly demanding forecasters, ensuring they meet strict compliance standards prior to deployment. The conditional diffusion models generate high-fidelity data that preserve complex spatial-temporal correlations and causal relationships between environmental factors and energy consumption. This tests whether forecasting models remain robust and accurate across highly stochastic daily and seasonal consumption dynamics.

The synthetic dataset serves as a diverse, multi-scale substrate for diagnosing downstream algorithmic vulnerabilities, structured around three evaluation axes:

- **Bias and Coverage Testing:** Downstream models can be systematically audited across explicitly decoupled conditioning layers to isolate localized performance decay. This framework isolates model **bias to seasonal data** (diurnal, weekly, and seasonal cycles), **regional data** (geographical shifts), **household characteristics** (floor areas, EPC classes, and asset adoption), and multivariate **weather patterns**. To formalize this, **coverage testing** evaluates algorithmic performance across distinct structural and behavioural clusters, ensuring model utility remains uniform even within sparse or atypical archetypes.
- **Robustness and Stress Testing:** The dataset evaluates model resilience under volatile or degraded conditions across two dimensions. **Temporal robustness** challenges forecasting and optimization networks with extended historical horizons and simulated climate anomalies to detect structural drift over time. Concurrently, **sampling frequency robustness** leverages the multi-scale training pool—spanning sub-minute, 15-minute, and hourly intervals—to verify that models maintain physical and structural consistency when resampled across disparate time resolutions.
- **Privacy Testing:** Because the generative pipeline relies entirely on probabilistic distributions and latent vectors, the dataset contains zero personal identifiers or sensitive corporate metadata. Third parties can train, tune, and evaluate complex energy intelligence models **without privacy concerns** or governance constraints. This setup provides a risk-free environment to confirm that downstream networks are capturing generalized behavioural and physical dynamics rather than memorizing or recovering private empirical profiles.

Taken together, the two evaluation roles are complementary: the source-anchored dataset establishes whether an AI system is valid against a realistic production environment, and the generative data establishes whether it remains robust beyond it — across unseen compositions, regions, and rare scenarios — so that compliance can be argued both in-distribution and out-of-distribution.

4.4.6 FAIR Principles for Data Sharing

Explain how the synthetic dataset is shared and managed according to FAIR principles to ensure it can be used effectively for AI testing and compliance purposes.

The openly publishable artefact in this contribution is the source-anchored re-synthesised dataset, so the FAIR commitments below are framed around its release; the generative method aligns with the same principles by relying only on publicly available datasets and by publishing its generation methods, and the data it produces inherits the same open, non-personal status.

Draft

Open-science release under the FAIR principles (Findable, Accessible, Interoperable, Reusable) is a deliverable-level commitment of the CERTAIN project rather than an afterthought. The dataset is conceived from the outset as a third-party-usable artefact accompanied by a future peer-reviewed companion data paper, and each FAIR sub-principle is mapped to a concrete operational decision in the synthesis and release pipeline.

Findable: The dataset will be published via a public repository that issues a persistent, version-pinned DOI for every release. Repository selection is still to be decided on and is constrained by capacity for the full release volume (estimated tens of GB of compressed Parquet), support for versioning, durability beyond the CERTAIN project lifetime, and machine-readable citation metadata in DataCite or equivalent schema. Candidate repositories include Zenodo, IEEE DataPort, Harvard Dataverse, and an institutional repository at the University of Applied Sciences St. Pölten. Discoverability is reinforced by the companion data paper, by indexed metadata records carrying domain keywords (synthetic data, energy community, smart meter, GDPR, Austria, EAG), and by cross-references in the CERTAIN project's public dissemination outputs.

Accessible: The dataset will be freely downloadable without registration, contractual restriction, or data-use agreement — a deliberate departure from the access regimes of comparable real datasets in the smart-meter space. The release format is open and standardised (Apache Parquet), accessible via a wide range of free tools (DuckDB, Pandas, Apache Arrow). The full schema documentation, including SQL DDL, OBIS code references, and field-level definitions, ships with the dataset and is also accessible independently via the public repository. Source code for the synthesis pipeline is hosted on a public GitHub repository, allowing third parties to inspect the methodology, audit the synthesis, and (subject to access to compatible source data) regenerate the dataset. Data lineage from input summary to published Parquet files is tracked via DVC.

Interoperable: The schema follows community standards and is explicitly aligned with regulatory framework. Measurement types use the official OBIS code system (1-1:1.9.0 G.01T, 1-1:2.9.0 G.01T, 1-1:2.9.0 P.01T, 1-1:2.9.0 G.03, 1-1:2.9.0 G.02), making the dataset directly interoperable with any system that consumes EAG-compliant smart-meter data. Timestamps are persisted in UTC with explicit timezone awareness and explicit period durations, eliminating the temporal-format ambiguities that plague comparable datasets. Geographic coordinates use a standard latitude/longitude representation. Identifier formats (metering-point numbers, participant numbers, community numbers) follow Austrian conventions. The Parquet column-typed schema is preserved across the relational mapping to SQL, so the dataset can be ingested into any conventional time-series back-end (PostgreSQL with TimescaleDB, DuckDB, or other) without transformation. Vocabularies and enumerations (energy direction, energy source, measurement type, participant type) are documented and stable.

Reusable: The dataset will be released under an open license (the specific license, e.g. CC BY 4.0, CC BY-SA 4.0, ODbL, or equivalent is still a decision to be made) that permits free redistribution, derivative works (including derivative datasets and derivative trained models), and downstream training without copyleft on resulting models, subject to attribution to the dataset and the companion data paper. Provenance is documented end-to-end: every published release is associated with a tagged synthesis-pipeline version, a fixed random seed, a complete configuration file capturing every parameter of every stage, and a DVC-tracked data lineage from input summary through cluster fits to published output. The companion data paper documents the methodology, the quality acceptance gates, the achieved per-cluster fidelity / utility / privacy metrics, the known limitations (single-source ecosystem, one-year temporal scope, no behind-the-meter device decomposition, exclusion of wind generation, thin coverage of multi-EEG cross-membership), and the future-extension axes (battery storage, EV charging, multi-year horizons, real-time streaming variants). Quality information is therefore explicit and machine-readable rather than implicit, allowing downstream users to make informed decisions about which clusters and which use cases the dataset is fit for. Ethical and legal positioning under the GDPR Recital 26 standard is explicitly documented (full re-synthesis from clustered structural and statistical features, with the source-to-synthetic mapping non-existent by construction), grounding the dataset's reusability in evaluation, compliance, and demonstrator contexts where pseudo-anonymised real data could not be used.

4.4.7 Data Protection (RoPA / DPIA) and Fairness

Following the FAIR commitments above, this subsection states the combined data-protection and fairness position for the energy contribution. Because the two synthesis regimes stand in different relations to personal data, the data-protection obligations apply asymmetrically across the two methods, whereas the fairness position is shared.

© 2024-2027 CERTAIN

Data protection (RoPA / DPIA): The obligations under the Records of Processing Activities (RoPA) and the Data Protection Impact Assessment (DPIA) attach specifically to the source-anchored (SPU) path, which processes real personal data from the EMPOWER ecosystem during source preparation. These obligations are discharged at the synthesis boundary: all personally identifiable information is irreversibly destroyed before any data crosses into the synthesis pipeline, the real-to-anonymised mapping is held only by the source pilot and never exposed, and every downstream artefact is by construction free of source PII, meeting the GDPR Recital 26 standard for anonymity. The generative (D4P) path uses only publicly available datasets and publicly available generation methods and processes no personal data; it therefore falls outside the scope of RoPA/DPIA obligations by construction. The published dataset that results from either path contains no personal data and is itself outside the scope of further processing records.

Fairness: Both methods support fairness and bias evaluation as a shared property of the contribution rather than of one method alone. The source-anchored dataset's explicit cluster structure (residential, commercial, and industrial consumers; PV and hydro generators) and per-cluster quality reporting expose heterogeneity that single-score datasets obscure, enabling tests of whether an AI system systematically under-performs on particular profiles; its known single-source coverage limit is disclosed so that fairness claims are bounded to the represented population. The generative method complements this by synthesising under-represented or absent profiles — alternative household compositions, technology-adoption patterns, regions, and countries — so that coverage gaps in any single source can be filled deliberately rather than left as silent bias. Together, bias can be both measured against a realistic in-distribution benchmark and mitigated by generating controllable coverage, which is the fairness capability the deliverable commits to.

4.4.8 Published Dataset: Metadata and Schema

The source-anchored dataset described in this section is not released together with this deliverable; it will be published as an openly available artefact in a future step of the project, once the outstanding release parameters (licence, repository, persistent identifier, and final temporal coverage) are fixed. The catalogue below specifies how the dataset will be described on publication: its descriptive identification, the provenance and source ecosystem the synthesis is grounded in, the scope and structure of the published artefact, the external reference data and its licensing, and the release conditions, followed by the field-level data schema. Fields whose final value is still an open release decision are shown as placeholders in red; they are intentional release parameters, not gaps in the schema design.

Dataset Metadata

Table 4-12 Identification

Field	Value
Title	Synthetic Energy Community Dataset
Description	A fully re-synthesised, openly publishable dataset of Austrian renewable energy communities (Energiegemeinschaften, EEG) at 15-minute resolution. It reproduces the structural and statistical properties of real community-level smart-meter data, namely participation factors, weighted consumption and generation, and community coverage and surplus, without any personal data, to support community-level energy research under open-science conditions.
Producers / responsible parties	Bürgerenergiegemeinschaft EMPOWER; USTP University of Applied Sciences St. Pölten, Data Intelligence Research Group
Project context	CERTAIN (Certification for Ethical and Regulatory Transparency in Artificial Intelligence); data foundation for the CERTAIN energy pilot.
Contact	Dipl.-Ing. Fabian Kovac, BSc University of Applied Sciences St. Pölten fabian.kovac@ustp.at
Persistent identifier (DOI)	Assigned by the archiving repository upon dataset publication
Version	Assigned at dataset publication

© 2024-2027 CERTAIN

Provenance and Source Foundation

Every published value is generated, not de-identified: there is no natural person behind any record. The synthesis is grounded in the real source ecosystem characterised below; the structural and statistical properties of that ecosystem are preserved, while the individual identities are absent rather than transformed.

Table 4-13 provenance and source foundation

Field	Value
Synthesis approach	Full re-synthesis (no pseudo-anonymisation). No personal data; no real metering point, name, address, or e-mail is reproduced.
Source ecosystem	EMPOWER citizen energy community ecosystem (real Austrian energy communities operating under a single, coherent EAG / network-operator / OBIS regulatory frame).
Source communities	Up to 28 energy communities.
Source members / metering points	3,425 members across 6,645 metering points (as of June 2026).
Source temporal coverage	From 01.01.2024 to 30.06.2026 spanning two and a half years of data. Coverage is not uniform across metering points. The communities grew over this period, so each smart-meter series begins when its metering point joined, and individual series lengths vary substantially.
Generation sources covered	Photovoltaic (PV) and small hydropower only. Wind is excluded: the source data contains only two wind metering points, below the k ≥ 5 threshold required by the dataset's privacy criteria.

Table 4-14 Published Dataset Scope and Structure

Field	Value
Temporal resolution	15-minute intervals, time-zone aware (UTC-persisted).
Temporal coverage (published)	Calendar year 2025 (35,040 quarter-hour intervals per metering point), finalised at publication
Spatial domain	Austria, where the catchment area is roughly Lower Austria (Niederösterreich). Addresses are real, existing addresses, each geographically consistent with its lat/lon coordinates. They bear no relation, however, to the real-world metering points or the persons behind them: the assignment of an address and its coordinates to a metering point is randomly sampled. Locations are drawn from within the source region, so that the weather dependency of the energy series is preserved.
Data format	Each relation is published in two formats: Parquet, which preserves column field types and is more performant and resource-efficient than CSV; and CSV, for simplicity and readability.
Energy schema tables	Entities: energy_community, participant, metering_point, energy_data. Relations: participant_metering_point, energy_community_participant, participation_factor_history. Full field-level specification in the Energy Data Schema below.
Reference schema tables	EPEX SPOT 1-Day Ahead Price AT; Weather data. Full field-level specification in the Reference Data Schema below.
Measurement semantics	OBIS-aligned measurement families (WMC, WMG, WSG, CC, CP); all energy values in kWh and non-negative. Participation factors are first-class, time-varying data (participation_factor_history). Each metering point belongs to exactly one participant for the lifetime of the dataset.
Record counts (published)	Fixed at publication; on the order of millions of metering-point quarter-hours (Section 4.4.2)

Draft

© 2024-2027 CERTAIN

External Reference Data

The weather reference data is not synthesised. It is sourced externally from GeoSphere Austria and integrated through a deliberate, dataset-specific preprocessing step so that it is structurally uniform with the energy data.

Table 4-15 Published dataset metadata: external reference data sources

Field	Value
Weather data source	GeoSphere Austria Data Hub, INCA analysis (resource inca-v1-1h-1km). The resource provides an hourly, 1 km gridded historical analysis.
Weather data licence	Creative Commons Attribution 4.0 International (CC BY 4.0). Attribution to GeoSphere Austria is carried through to this dataset's release.
Preprocessing	The source INCA data is delivered as hourly, wide-format grids. We deliberately apply our own preprocessing: the data is pivoted into long format (one row per coordinate, interval, and observation_type) so that it shares the same period_begin / period_duration / observation_type / observation_value shape as the energy data and links to communities and objects via coordinates. Temporal alignment to the 15-minute energy grid is part of this preprocessing.
Market price data source	EPEX SPOT 1-Day Ahead Price (Austria), stored in cent/kWh at 15-minute resolution.

Table 4-16 Legal Basis and Release Conditions

Field	Value
Data protection status	No personal data within the meaning of the GDPR. Because every record is fully synthesised (not pseudonymised under Art. 4(5)), no natural person is identifiable, consistent with the anonymity standard of GDPR Recital 26.
Regulatory alignment	Austrian Erneuerbaren-Ausbau-Gesetz (EAG, BGBl 150/2021): EAG-compliant community entities, OBIS-coded measurements, participation-factor mechanics.
Dataset licence	An open licence permitting redistribution and derivative works (CC BY 4.0 or equivalent), finalised at publication
Repository / access point	Zenodo, IEEE DataPort, Harvard Dataverse, or the St. Pölten institutional repository (selection at publication)
Release date	Following this deliverable, at dataset publication
Recommended citation	Provided with the DOI at publication

Data Schema

The schema is the contract between the dataset and its consumers. It is specified at the field level, with types, units, and OBIS codes, for both the energy data and the external reference data.

Energy Data Schema

The field-level schema below is the canonical specification of the energy data. It comprises the schema's entities, which are the independent objects, each with its own identity and surrogate key, together with the relations that link them via foreign keys and the enumerations they reference. The entity-relationship diagram is shown first, followed by the field-level tables.

Draft

© 2024-2027 CERTAIN

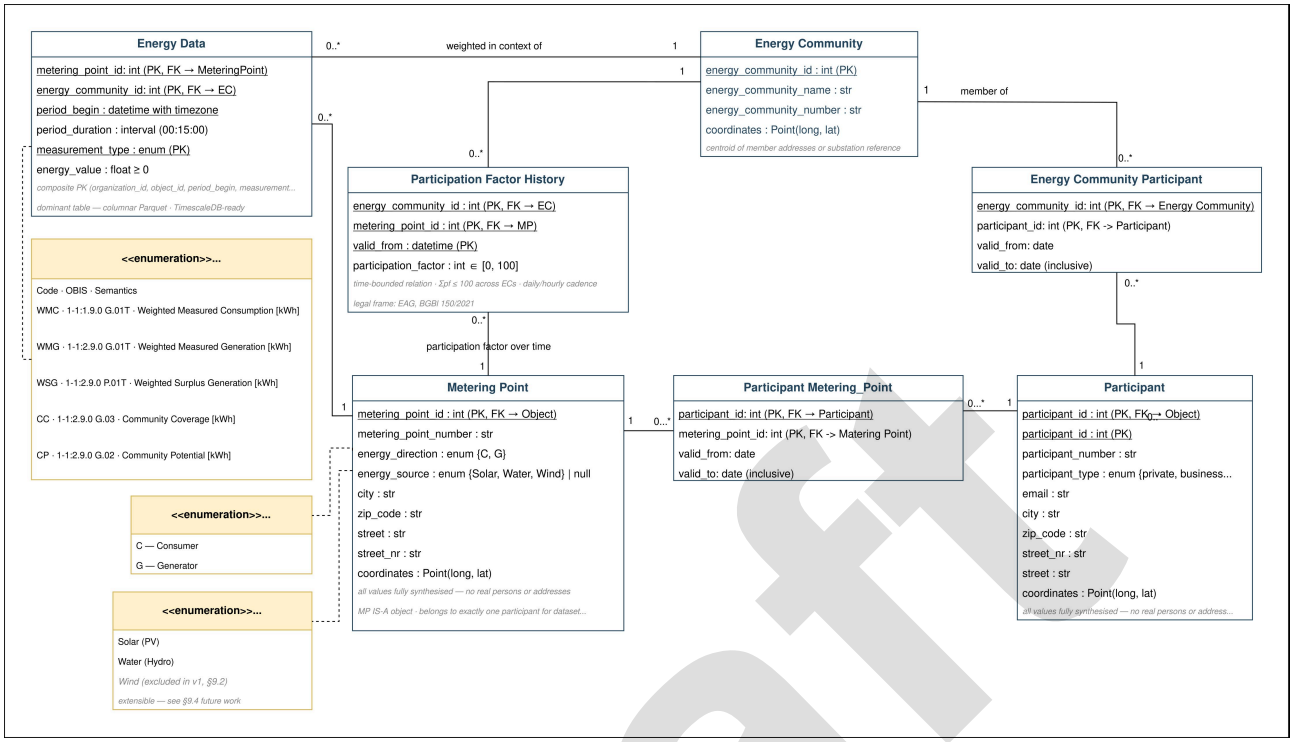

Figure 4-13 Energy data model showing the entities (energy_community, participant, metering_point, energy_data), their relations, and the measurement-type, energy-source, and energy-direction enumerations.

energy_community

Static descriptive data of an EEG.

Table 4-17 Energy data schema: energy_community entity

Field	Type	Description
energy_community_id	int (PK)	Surrogate identifier
energy_community_name	str	Synthetic community name
energy_community_number	str	Public community number
coordinates	Point(long, lat)	Single-point geographic representation of the EEG

participant

Members of an energy community.

Table 4-18 Energy data schema: participant entity

Field	Type	Description
participant_id	int (PK)	Surrogate identifier
participant_number	str	Public participant number
participant_type	str	Member type (private, business, etc.)
email	str	Synthetic email under which the participant is registered
participant_name	str	Synthetic full name of the participant

Draft

street, street_nr, zip_code, city	str	Synthetic Austrian address
coordinates	Point(long, lat)	Geographic representation of the participant's address

metering_point

The smart-meter unit attached to a generation or consumption point.

Table 4-19 Energy data schema: metering_point entity

Field	Type	Description
metering_point_id	int (PK)	Surrogate identifier
metering_point_number	str	Structurally valid Austrian metering-point number, but synthetic (no real meter)
energy_direction	enum {C, G}	C = Consumer, G = Generator
energy_source	enum {Solar, Water, Wind}	For generators only; null for consumers
street, street_nr, zip_code, city	str	Synthetic Austrian address
coordinates	Point(long, lat)	Geographic representation of the metering point's address

energy_data

The actual time-series data, in 15-minute resolution. All values are non-negative.

Table 4-20 Energy data schema: energy_data entity

Field	Type	Description
metering_point_id	int (PK, FK)	Reference to metering point
energy_community_id	int (PK, FK)	Reference to energy community
period_begin	datetime with tz (UTC-persisted)	Start of the 15-minute interval
period_duration	interval	00:15:00
energy_value	float ≥ 0	Measurement value (kWh)
measurement_type	enum (see enumeration below)	OBIS-aligned measurement family

participant_metering_point

Ownership mapping between participants and metering points.

Table 4-21 Energy data schema: participant_metering_point relation

Field	Type	Description
participant_id	int (PK, FK)	Reference to participant
metering_point_id	int (PK, FK)	Reference to metering point
valid_from	date	Start date of the ownership
valid_to	date	Inclusive end date of the ownership

Draft

energy_community_participant

Membership mapping between energy communities and participants.

Table 4-22 Energy data schema: energy_community_participant relation

Field	Type	Description
energy_community_id	int (PK, FK)	Reference to energy community
participant_id	int (PK, FK)	Reference to participant
valid_from	date	Start date of the membership
valid_to	date	Inclusive end date of the membership

participation_factor_history

The time-bounded relation linking a metering point to an energy community.

Table 4-23 Energy data schema: participation_factor_history relation

Field	Type	Description
energy_community_id	int (PK, FK)	Reference to energy community
metering_point_id	int (PK, FK)	Reference to metering point
valid_from	datetime	Start of the schedule window
participation_factor	int ∈ [1, 100]	Percentage participation in the community at this time

Energy data measurement type enumeration

The enumeration of measurement families the energy_data table can carry, with their official OBIS codes.

Table 4-24 Energy data schema: measurement type enumeration (OBIS codes)

Code	Name	Semantics	Unit	OBIS
WMC	Weighted Measured Consumption	Consumption weighted by participation factor	kWh	1-1:1.9.0 G.01T
WMG	Weighted Measured Generation	Generation weighted by participation factor	kWh	1-1:2.9.0 G.01T
WSG	Weighted Surplus Generation	Residual surplus generation after community use	kWh	1-1:2.9.0 P.01T
CC	Community Coverage	Self-coverage within the community	kWh	1-1:2.9.0 G.03
CP	Community Potential	Community share potential	kWh	1-1:2.9.0 G.02

Reference Data Schema

The field-level schema below is the canonical specification of the reference data and the source of truth that the companion data paper will inherit.

Draft

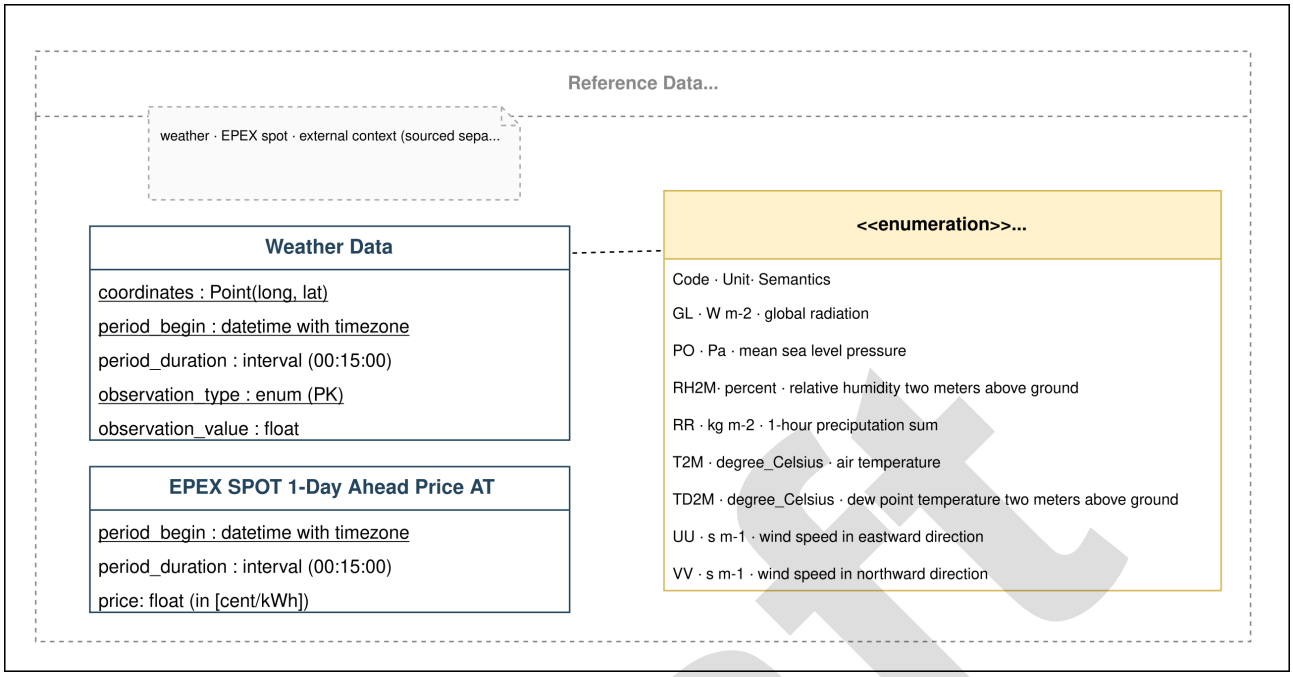

Figure 4-14 Reference data model showing the weather observations (GeoSphere INCA) and the EPEX SPOT 1-day-ahead Austrian market price, together with the weather observation-type enumeration.

EPEX SPOT 1-Day Ahead Price AT

The historical time-series of the EPEX SPOT 1-Day Ahead Price in Austria.

Table 4-25 Reference data schema: EPEX SPOT 1-Day Ahead Price AT

Field	Type	Description
period_begin	datetime with timezone (PK)	Start of the 15-minute interval
period_duration	interval	00:15:00
price	float	Spot-market price in cent/kWh

Weather data

The historical time-series of the weather data.

Table 4-26 Reference data schema: weather data (GeoSphere Austria INCA)

Field	Type	Description
coordinates	Point(long, lat) (PK)	Geographic representation of the weather data for linkage to energy community and objects.
period_begin	datetime with timezone (PK)	Start of the 15-minute interval
period_duration	interval	00:15:00
observation_type	enum (PK)	Type of the weather observation
observation_value	float	Actual value of the weather observation

Draft

Weather data observation type enumeration

The enumeration of observation types the weather_data table can carry.

Table 4-27 Reference data schema: weather observation type enumeration

Code	Semantics	Unit
GL	global radiation	W m-2
PO	mean sea level pressure	Pa
RH2M	relative humidity two meters above ground	percent
RR	1-hour precipitation sum	kg m-2
T2M	air temperature	degree_Celsius
TD2M	dew point temperature two meters above ground	degree_Celsius
UU	wind speed in eastward direction	s m-1
VV	wind speed in northward direction	s m-1

4.4.8.1 Generative Synthetic Dataset from Public Priors

In addition to the core datasets described in this deliverable, an Out-of-the-Distribution (OOD) dataset will be developed and integrated at a later stage of the project. This dataset will serve as a supplementary data source, providing more mature, realistic, and diverse scenarios. The OOD dataset will support the assessment of model robustness, and generalization capabilities. The following catalog presents the key features that are planned to be generated and included within this dataset.

Table 4-28 Feature catalogue of the planned out-of-distribution generative energy dataset

Feature	Explanation	Unit / Format
temperature	Air temperature. Influences heating and cooling demand.	°C (signed)
radiation_direct_horizontal	Direct solar radiation reaching a horizontal surface. Higher values indicate stronger sunlight.	W/m² (≥0, log1p)
radiation_diffuse_horizontal	Solar radiation scattered by clouds and the atmosphere.	W/m² (≥0, log1p)
solar_generation_actual	Actual solar electricity generation measured in the grid.	MW (≥0, log1p)
wind_onshore_generation_actual	Actual electricity generation from onshore wind farms.	MW (≥0, log1p)
household_consumption_MW	Electricity consumption measured at the household level.	MW (≥0)
grid_total_load_MW	Total electricity demand on the grid.	MW (≥0)

Draft

© 2024-2027 CERTAIN

grid_total_generation_MW	Total electricity generation supplied to the grid.	MW (≥0)
hour	Hour of the day.	0–23
day_of_week	Day of the week.	0–6
month	Month of the year.	1–12
day_of_year	Position of the day within the year.	1–365/366
is_weekend	Indicates whether the day is Saturday or Sunday.	Binary
is_holiday	Indicates whether the date is a public holiday.	Binary
season_idx	Represents the season (e.g., winter, spring, summer, autumn).	Categorical index
doy_sin / doy_cos	Cyclical encoding of day of year to capture annual seasonality.	Encoded
dow_sin / dow_cos	Cyclical encoding of day of week.	Encoded
hour_sin / hour_cos	Cyclical encoding of hour to preserve daily periodicity.	Encoded
country	Country identifier encoded as country_{CC}.	One-hot encoded
region	Region identifier encoded as region_{name}.	One-hot encoded
epc_class	Energy Performance Certificate rating (A–G).	One-hot encoded
dwelling_type	Type of residence (e.g., apartment, detached house, semi-detached).	One-hot encoded

Draft

© 2024-2027 CERTAIN

4.5 TEXT - PUBLIC DELIBERATION SYNTHETIC DATASETS GENERATION COMPONENT

The Public Deliberation Synthetic Datasets Generation Component (https://github.com/CERTAIN-Project/Synthetic-Data-Generation-Component) developed by NVCR is a fully functional system designed to generate realistic public deliberation datasets through AI-driven agent simulation. The system creates synthetic online discussion forums populated by diverse AI agents, enabling research and development in computational social science, democratic deliberation platform design, and natural language processing without relying on real-world public data. In the sections below, we present the component in detail, including its characteristics and methodology.

4.5.1 Domain Description

The Public Deliberation Synthetic Datasets Generation Component operates within the domain of online public deliberation and civic discourse. It is designed to simulate the kinds of conversations that take place on civic technology platforms and online forums where citizens engage with topics of public concern, ranging from local policy issues to broader socio-political questions. The system produces synthetic discussion threads that replicate the structural and linguistic characteristics of these environments, making the resulting data suitable for a wide range of AI research and development applications.

The primary AI applications for which this component provides test data include the development and evaluation of natural language processing pipelines for discourse analysis, the training and benchmarking of content moderation systems, the testing of bias detection algorithms across diverse participant profiles, and the assessment of dialogue understanding models in multi-turn conversational settings. More broadly, the system supports research into computational social science, the design and iterative improvement of deliberation platforms, and the evaluation of tools intended to support democratic participation online.

A key strength of the component is that it enables all of these research activities without dependence on real-world public discussion data. Researchers can study how different demographic groups engage with complex topics, test how AI systems handle incomplete or degraded inputs, and evaluate whether their tools perform equitably across the simulated population, all within a controlled environment that carries no risk of exposing personal information or breaching research ethics requirements.

4.5.2 Source Data Characteristics

The Public Deliberation Synthetic Datasets Generation Component does not draw upon any real-world dataset as its basis. This is a deliberate architectural decision, reflecting the consortium's commitment to privacy-by-design principles. No real conversations from actual individuals are collected, processed, stored, or referenced at any stage of the generation pipeline. Every element of every dataset including all post content, participant identities, and demographic attributes is generated from scratch.

Each dataset takes the form of a structured online discussion thread. Posts and replies are organised chronologically and attributed to one of 23 fictitious participants. The dataset structure conforms to a standardised JSON schema in which each record contains three core elements: a unique post identifier (a UUID compliant with RFC 4122), the text content of the post, and a metadata object containing the participant's demographic attributes.

For identified participants, the metadata object includes six fields which are fully adaptable: first name, last name, age, date of birth, nationality, and occupational background. For anonymous participants, all metadata fields are set to null, reflecting the absence of identifying information in real-world anonymous contributions. The demographic attributes of identified participants are procedurally generated to represent a realistic distribution of diversity. To produce the datasets to be used in the framework of the project we made specific selections for each metadata field. Ages range from 25 to 70 years, nationalities are drawn from five countries (Greece, Cyprus, Bulgaria, Romania, and Albania), and occupational backgrounds span ten categories including engineer, teacher, doctor, manager, student, consultant, developer, designer, sales professional,

Draft

© 2024-2027 CERTAIN

and human resources practitioner. All participant names reflect the Greek and Eastern European cultural context of the target region.

In terms of temporal structure, each dataset represents a simulated discussion unfolding over a configurable period, with a default simulation horizon of 1.5 minutes of simulated time. Posts are logged chronologically in the order in which they are generated, and reply relationships between posts are tracked to preserve the threaded structure of the conversation. A typical simulation produces between 20 and 50 posts, depending on the activity parameters configured for that run. The dataset thus captures both the content and the conversational dynamics of a deliberative exchange, including the distribution of top-level posts (approximately 55% of all contributions) and replies to existing messages (approximately 45%).

The system produces two variants of each dataset; a clean version and a dirty version, which are described in detail in the next section. This paired design is important because it allows researchers to evaluate AI systems both under ideal data conditions and under realistic data quality challenges. The clean dataset provides a reliable baseline for model development, training, and initial validation, while the dirty dataset simulates common issues such as missing or inconsistent metadata, shuffled participant attributes, and truncated text. Together, these variants support robustness testing, data-cleaning research, and comparative assessment of system performance when operating on both pristine and degraded inputs.

4.5.3 Synthetic Data Generation Methodology

The datasets are generated through the coordinated operation of three distinct techniques, each of which addresses a different dimension of realistic deliberation simulation. The first technique is an agent-based simulation. The system instantiates 23 artificial agents, each representing a different type of discussion participant. Fifteen of these are standard identified participants, carrying full demographic profiles and exhibiting the behavioural characteristics of the citizen archetype: natural, conversational language, moderate and varied posting frequency, opinion-based reasoning with occasional references to personal experience, and a balanced mix of initiating new threads and responding to existing ones. Two further identified participants — named Vasilis Kritikos and Despina Skepsi — are configured with enhanced personality markers that influence the tone and content of their generated posts whilst remaining grounded in the citizen archetype. The remaining six participants are anonymous, contributing posts and replies without any associated demographic information. The simulation runs through a series of discrete time steps, and at each step the system determines, on the basis of configurable parameters, which agents post and whether each contribution takes the form of a new top-level comment or a reply to an existing message.

The second technique deployed is AI-driven natural language generation. When an agent is scheduled to post, the system invokes an LLM to generate the post content. The generated text is conditioned on three contextual inputs: the topic under discussion, the personality archetype of the posting agent, and the full history of the conversation up to that point. This approach ensures that each post is topically coherent, stylistically appropriate to the participant type, and responsive to what has already been said in the thread. No text is extracted or adapted from any external source; all content is generated fresh on each invocation.

The third technique is procedural demographic synthesis. Participant names, ages, nationalities, and occupational backgrounds are generated algorithmically according to configured distributions, ensuring realistic population diversity without any correspondence to real individuals.

Two constraints are applied consistently across the generation process. The first is reproducibility: the system is deterministic given a fixed random seed, meaning that any dataset can be regenerated exactly by re-running the simulation with the same seed and configuration parameters. This property is essential for scientific reproducibility and for enabling other researchers to verify published results. The second constraint is the paired clean/dirty structure of outputs.

The clean dataset contains the full, unmodified output of the simulation, with complete and accurate metadata for all identified participants and untruncated post content. It is suited to primary research, model training, and the development of baseline algorithms. The dirty dataset mirrors the clean dataset in content but incorporates systematic, controlled degradation to simulate the kinds of data quality issues that arise in real-world field data collection. Specifically, 30% of identified participant records have their metadata shuffled (assigning a different participant's demographic profile to the post), 50% have one or two metadata fields randomly nullified to simulate incomplete data, and 40% have their post text truncated to simulate cut-off messages. Anonymous participant records in the dirty dataset retain their null metadata consistently. This paired approach allows

© 2024-2027 CERTAIN

researchers to directly compare algorithm performance under ideal and degraded data conditions, and to develop and evaluate data cleaning and imputation strategies, without the need to manually corrupt datasets.

4.5.4 Intended Use for Compliance Testing

The Public Deliberation Synthetic Datasets Generation Component is specifically designed to serve as a testing infrastructure for AI systems that operate on online discussion data. Because the datasets are entirely synthetic, they can be used in testing regimes that would be legally or ethically problematic with real data, shared freely across institutions and research groups, and adapted or modified without restriction, subject to the attribution requirements of the licence. The component supports five principal categories of compliance and validation testing.

1. **Bias testing:** The controlled demographic composition of the simulated population allows researchers to construct scenarios in which specific participant characteristics — nationality, age, occupational background — are varied systematically, enabling rigorous evaluation of whether AI systems treat different groups equitably.
2. **Robustness testing**, for which the dirty dataset is particularly well suited. By evaluating system performance on data containing missing fields, shuffled metadata, and truncated text, researchers can assess whether their tools degrade gracefully under realistic field conditions or exhibit fragile failure modes.
3. **Stress testing with rare or unusual scenarios**: The configurable nature of the simulation makes it straightforward to generate edge cases, atypical participation patterns, or unusually polarised discussions that might be difficult to find in real-world corpora but are important for validating system resilience.
4. **Compliance verification**, whereby tools designed to ensure regulatory compliance for instance, content moderation systems operating under the EU AI Act can be tested against the synthetic data to confirm that they behave as specified before deployment with real users.
5. **Training of new AI models:** The synthetic datasets can be used as training corpora for dialogue systems, content classifiers, and deliberation analysis models, with no privacy constraints on the volume or distribution of data generated.

The validity of the synthetic data for these purposes rests on the ecological plausibility of the generated conversations. The combination of agent-based simulation and context-aware language generation produces discussions that exhibit realistic conversational structure, appropriate stylistic variation across participant types, and logical coherence across turns. The paired clean/dirty structure further extends the utility of the datasets by covering the full range of data quality conditions that AI systems are likely to encounter in practice.

4.5.5 FAIR Principles for Data Sharing

All datasets produced by the Public Deliberation Synthetic Datasets Generation Component are managed and distributed in accordance with the FAIR principles which constitute the standard framework for responsible scientific data sharing in the European research infrastructure.

Findable: Each dataset shall have a persistent Digital Object Identifier (DOI) assigned at the point of export, providing a permanent and unambiguous reference that does not change over time. Individual posts within each dataset carry RFC 4122 compliant UUIDs, enabling granular citation and traceability at the record level. Datasets shall be registered major scientific repositories like Zenodo, the Open Science Framework (OSF), and DANS (Data Archiving and Networked Services), ensuring broad discoverability across the European and international research community. Standardised Dublin Core metadata shall be included with each dataset, describing the topic, generation date, version number, and provenance, in a machine-readable format that supports automated indexing and search.

Accessible: Clean datasets shall be published under a Creative Commons Attribution 4.0 International (CC-BY 4.0) licence and be available free of charge through multiple channels, like GitHub, Zenodo, and direct HTTP download, without the need for authentication or specialised software. The terms of use shall be clearly stated with each dataset. Dirty datasets and intermediate processing logs, which require more controlled distribution, shall be accessible to consortium members via authenticated cloud storage managed through

© 2024-2027 CERTAIN

MinIO, with full audit logging of access events. The licence and access conditions are designed to ensure that the data remains permanently available and that its terms are unambiguous to all potential users.

Interoperable: All datasets are provided in JSON format, which is natively supported by every major programming language and data analysis environment. The data conforms to a published JSON Schema specification, and metadata is provided in JSON-LD format to support semantic interoperability. The schema uses standardised vocabularies: nationality fields use ISO 3166-1 country names, occupational backgrounds are drawn from an enumerated set of categories, and post identifiers conform to RFC 4122. Data transformation support is provided for the most common analytical environments, including CSV export for spreadsheet and statistical software (Excel, SPSS), direct import via `json_normalize()` for Python-based data science workflows, `jsonlite::fromJSON()` compatibility for R, Parquet format for Apache Spark and large-scale data processing, and SQL insert templates for PostgreSQL and SQLite.

Reusable: The CC-BY 4.0 licence permits a broad range of uses without restriction beyond attribution, including academic research and publication, university teaching, commercial platform testing, training of machine learning models, and integration into products and services. Derivative datasets are permitted provided that modifications are clearly marked and the same licence is applied. The datasets include complete provenance documentation (simulation parameters, random seeds, configuration fingerprints, and output hashes) enabling any researcher to understand exactly how a dataset was produced and to replicate it precisely. Datasets shall be archived with Zenodo under the LOCKSS (Lots of Copies Keep Stuff Safe) preservation model, ensuring availability for at least 20 years. All dataset versions shall be maintained in GitHub with full version history. The combination of open licensing, complete documentation, persistent identifiers, and long-term archival commitment ensures that the datasets will remain useful, understandable, and reproducible for the foreseeable future.

4.6 HUMAN RESOURCES

4.6.1 Domain Description

The Human Resources (HR) domain (Pilot 4) within the CERTAIN project, focuses on enabling data-driven human resources management processes, supporting recruitment, workforce analytics, and policymaking using AI-driven tools and services. The pilot involves the exchange and processing of heterogeneous data, including job opportunities, candidate/freelance experts' profiles (including Curriculum Vitae (CVs)), skills information, and organizational data, to enable improved matching, decision-making, and labour market and employment insights. In this context, data sharing serves as a key enabler for cross-organisational collaboration between different stakeholders in the human resources ecosystem, while ensuring that personal and organisational data is handled in a secure and controlled manner.

AI applications in this domain include:

- Federated job matching and candidate recommendation
- Skills extraction and competency profiling
- AI-assisted recruitment decision support
- Workforce analytics, labour market and employment trend analysis

A key feature of this domain is the requirement to operate under strict legal and ethical constraints, particularly those related to personal data protection, fairness, and transparency. As such, the HR use case focuses on compliance-aware AI deployment, rather than data-intensive model development in isolation.

Within CERTAIN, the HR domain serves as a **pilot for testing AI systems in a federated dataspace**, where data remains under the full control and governance of the original data holders and is accessed through controlled, interoperable mechanisms.

Draft

© 2024-2027 CERTAIN

4.6.2 Source Data Characteristics

The HR domain relies on real-world data structures typically used in recruitment and workforce management, while ensuring that no sensitive or identifiable data is exposed within the project.

The data involved includes:

- CVs and professional profiles (semi-structured and unstructured text)
- Job descriptions and Terms of Reference for specific expertise provision
- Skills and competency frameworks
- Recruitment workflow data (e.g. candidate selection stages, matching outputs)

Key characteristics of the data:

- **Dataset type**: Mixed (textual, tabular, and relational data)
- **Feature types**:
  - Textual (CV narratives, job descriptions)
  - Categorical (job roles, industries, skills categories)
  - Numerical (years of experience, duration of employment)
  - Relational (career trajectories, skill-job mappings)
- **Number of features**: Not fixed, but typically includes dozens of attributes per profile (e.g. skills, education, experience, certifications, languages)
- **Temporal aspects**: Strong temporal dimension, including:
  - Career progression over time
  - Employment history
  - Recruitment cycles and hiring processes
- **Data governance model**: Data is not centrally stored but remains within **federated systems**, with access mediated through secure connectors and identity management mechanisms.

Sensitive personal data is not disclosed or shared within the testing environment. The methodology should also define data governance controls for HR data, including role-based access, purpose-specific processing permissions, logging of access and queries, retention periods for source data, matching outputs and audit records, and deletion or anonymisation procedures after the relevant testing or compliance purpose has been fulfilled.

Where necessary, data is:

- anonymised or pseudonymised
- processed under strict access controls
- accessed only through compliance-aware workflows

The data exchange model within the CERTAIN Human Resources Data Space enables interactions between HR actors, AI systems, and end users in recruitment and workforce analytics processes. Recruiters and project managers provide job opportunities, candidate data, and recruitment metadata, which are consumed by AI systems to generate professionals matching results, skill insights, and analytics outputs. These outputs are fed back to support hiring decisions and process optimisation, while job seekers/freelance experts contribute CV data and receive personalised recommendations based on their professional experience and core expertise. Data sharing follows a controlled approach, where personal and organisational data are exchanged under policy enforcement and aggregated or anonymized outputs are shared for broader use, including policymaking and labour market analysis.

As synthetic data generation is not the primary method used in the HR domain, the methodology relies on controlled and federated processing of real-world data structures, anonymisation or pseudonymisation where appropriate, simulation of federated interactions, and scenario-based workflow testing. This allows the HR pilot to test AI-supported recruitment and workforce analytics under realistic legal, technical and organisational constraints without exposing raw personal data or requiring centralised datasets.

Draft

© 2024-2027 CERTAIN

4.6.3 Intended Use for Compliance Testing

In the HR domain, the dataset and infrastructure are used to evaluate AI systems under **realistic operational and regulatory conditions**, rather than to benchmark performance on synthetic datasets.

The intended uses include:

- **Fairness and bias testing**: Assessing whether AI-driven recruitment tools produce discriminatory outcomes across different groups
- **Robustness testing**: Evaluating system performance across diverse candidate profiles, job types, and recruitment scenarios
- **Federated system validation**: Testing the ability of AI tools to operate across distributed datasets without centralised access
- **Transparency and explainability testing**: Assessing whether AI-generated recommendations can be interpreted and justified
- **Compliance validation**: Verifying alignment with:
  - GDPR requirements (data minimisation, participants consent, lawful processing)
  - auditability and traceability requirements
  - emerging AI regulatory standards
- **Stress testing of workflows:** Simulating complex or edge-case recruitment scenarios (e.g. incomplete profiles, cross-platform searches, high-volume matching requests)

This approach ensures that AI systems are tested in a way that reflects real-world constraints, including legal, technical and organisational factors.

4.6.4 FAIR Principles for Data Sharing

The HR domain follows FAIR principles, adapted to a federated and compliance-sensitive environment:

- **Findable**: Data assets and services are indexed within the federated data space catalogue using standardised metadata, enabling discoverability without exposing personal data.
- **Accessible**: Access to data is controlled through:
  - authentication and authorisation mechanisms
  - role-based permissions
  - consent management systems

Data is not directly shared, but accessed through **secure federated queries**

- **Interoperable**: Data is structured using:
  - standardised schemas for HR data (e.g. skills, profiles, job descriptions)
  - semantic interoperability mechanisms
  - APIs and connectors enabling cross-platform interaction
- **Reusable**: Reusability is ensured through:
  - clear documentation of data structures and processing rules
  - traceability of data usage and transformations
  - compliance with legal and ethical constraints

In addition, the HR domain emphasises:

- **data sovereignty**, ensuring that data remains under the control of its original data owner
- **auditability**, enabling verification of how data is accessed and used
- **compliance-by-design**, integrating legal requirements into the data lifecycle.

For HR data, FAIR implementation should not be interpreted as open data sharing. Findability and reusability must remain subject to access control, purpose limitation, data minimisation, data-subject rights, confidentiality and the governance rules of the federated data space.

Draft

5 CONCLUSIONS

This deliverable had two objectives: to record the way in which the CERTAIN partners produced synthetic test data in their individual fields, and to combine those methods into a single methodology that could be used in all of them.

The various domains did not settle on a single approach, and this should be seen as a result rather than a drawback. The effectiveness of different methods depends on the particular data involved. In the case of the finance domain, counterfactuals based on financial time series were employed in order to examine and explain a profitability model. The health domain divided into two branches: on one side came the multimodal digital-phenotyping features from the Syn.Mind and SMILE cohorts, and on the other the radar breathing signals with apnea and motion artefacts having been injected. In the biometrics area, one partner created synthetic face identities using a GAN-and-diffusion process and demonstrated that these were anonymous with respect to the training data, while a second partner looked at and evaluated embedding-space generators for the same objective. The energy domain carried out two complementary methods simultaneously: a re-synthesis of a real Austrian energy community based on source anchoring, and a generative model trained solely on public data that can be applied to other countries and conditions. In the text domain, deliberation forums were constructed from simulated agents. Human resources is different from the other domains since the main limitation there is legal access to actual CVs rather than a lack of data, so very little synthetic data was generated, and the contribution consisted in stating this clearly and working within a federated setup instead.

The element that is common to all the different areas is the way in which the work was carried out. In each case, a real or reference dataset was used, a generation method appropriate to that dataset was selected, and the output was then assessed using the same three questions: whether it is faithful to the real data, whether it is useful for the downstream task, and whether it protects the people associated with the source dataset. These three criteria—fidelity, utility and privacy—are the core of the common methodology and are directly incorporated into the trustworthiness testing performed in Task 5.2.

DRAFT

6 REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, 2014.

[2] N. C. Abay, Y. Zhou, M. Kantarcioglu, B. Thuraisingham and L. Sweeney, "Privacy Preserving Synthetic Data Release Using Deep Learning," 2018.

[3] J. Voas and R. Kuhn, "Artificial Intelligence Risk Management," 2021.

[4] E. Commission, "Proposal for a Regulation on Artificial Intelligence (Artificial Intelligence Act)," 2021.

[5] OECD, "OECD Framework for the Classification of AI Systems," 2021.

[6] E. Commission, "Ethics Guidelines for Trustworthy AI," 2019.

[7] K. El Emam, S. Rodgers and B. Malin, "A systematic review of re-identification attacks on health data," *PLOS ONE,* 2020.

[8] J. Sanz-Cruzado, R. McCreadie, N. Droukas, C. Macdonald and I. Ounis, "Investors Are (Not) Always Right: A Comparison of Transaction-Based and Profitability-Based Metrics for Financial Asset Recommendations," *ACM Transactions on Information Systems,* pp. Article No.: 51, Pages 1 - 56, 2026.

[9] D. Thumm and . L. O. Mijares, "Towards Causal Market Simulators," in *ICAIF Workshop on Rethinking Financial Time-Series*, 2025.

[10] X. Wang, Y. Li, H. Zhang and Y. Shan, "Towards Real-World Blind Face Restoration with Generative Facial Prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[11] R. Marchesi, . N. Micheletti, N. I.-H. Kuo, S. Barbieri, G. Jurman and V. Osmani, "Generative AI mitigates representation bias and improves model fairness through synthetic health data," in *PLOS COMPUTATIONAL BIOLOGY*, 2025.

[12] F. Mandreoli, D. Ferrari , V. Guidetti, F. Motta and F. Mott, "Real-world data mining meets clinical practice: Research challenges and perspective," *Frontiers in Big Data,* 2022.

[13] O. D. Institute, "The Data Ethics Canvas," 2019.

[14] J. Ho and T. Salimans, "Classifier-Free Diffusion Guidance," 2022.

[15] A. Sharma, R. Mothilal and C. Tan, "Diverse counterfactual explanations (DiCE) for ML," [Online]. Available: https://interpret.ml/DiCE/.

[16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Draft

[17] X.-J. Mao, C. Shen and Y.-B. Yang, “Image Restoration Using Convolutional Auto-Encoders with Symmetric Skip Connections,” 2016.

[18] S. Li, J. Fu, K. Liu, W. Wang and R. Qi, “ID3: Identity-Preserving Diffusion Model for Face Recognition Data Generation,” in *Advances in Neural Information Processing Systems*, 2024.

[19] B. F. Labs, S. Batifol, A. Blattmann, F. Boesel, S. Consul, C. Diagne and L. Smith, “FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space,” 2025.

[20] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes,” 2013.

[21] M. Kim, F. Liu, A. Jain and X. Liu, “DCFace: Synthetic Face Generation with Dual Condition Diffusion Model,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[22] T. Karras, T. Aila, S. Laine and J. Lehtinen, “Progressive Growing of GANs for Improved Quality, Stability, and Variation,” 2017.

[23] T. Karras, S. Laine and T. Aila, “A Style-Based Generator Architecture for Generative Adversarial Networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[24] S. Kang, Y. Lee, Y.-H. Lim, H.-K. Park, S. H. Cho and S. H. Cho, “Non-Contact Diagnosis of Obstructive Sleep Apnea Using Impulse-Radio Ultra-Wideband Radar,” *Sensors,* vol. 21, no. 7, 2020.

[25] R. Ismayilov, D. Sero and L. Spreeuwers, “FLUXSynID: A Framework for Identity-Controlled Synthetic Face Generation with Document and Live Images,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

[26] Z. Zhu, G. Huang, . J. Deng, Y. Ye2, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du and J. Zhou, “WebFace260M: A Benchmark Unveiling the Power of Million-Scale Deep Face Recognition,” in *Conference on Computer Vision and Pattern Recognition (CVPR). IEEE*, 2021.

[27] Y. Guo, L. Zhang, Y. Hu, . X. He and J. Gao, “MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition,” in *European conference on computer vision. Cham: Springer International Publishing*, 2016.

[28] J. Deng, J. Guo, N. Xue and S. Zafeiriou, “ArcFace: Additive Angular Margin Loss for Deep Face Recognition,” in *IEEE/CVF conference on computer vision and pattern recognition.*, 2019.

[29] X. An, X. Zhu, Y. Xiao, L. Wu, M. Zhang, Y. Gao, B. Qin, D. Zhang, Y. Fu and J. Deng, “Partial FC: Training 10 Million Identities on a Single Machine,” in *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). IEEE*, 2021.

[30] R. Gross, I. Matthews, J. Cohn, T. Kanade and S. Baker, “Multi-PIE,” in *Int Conf Autom Face Gesture Recognit.* , 2010.

[31] D. Yi, . Z. Lei, S. Liao and S. Z. Li, “Learning Face Representation from Scratch,” 2014.

[32] M. Wang, W. Deng, J. Hu, X. Tao and Y. Huang, “Racial Faces in-the-Wild: Reducing Racial Bias by Information Maximization Adaptation Network,” in *EEE/CVF International Conference on Computer Vision (ICCV). IEEE.*, 2019.

Draft

© 2024-2027 CERTAIN

[33] G. B. Huang, . M. Mattar, T. Berg and . E. Learned-Miller, “Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments,” in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition.*, 2008.

[34] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa and D. W. Jacobs, “Frontal to Profile Face Verification in the Wild,” in *IEEE winter conference on applications of computer vision (WACV). IEEE,*, 2016.

[35] T. Zheng and W. Deng, “Cross-Pose LFW: A Database for Studying Cross-Pose Face Recognition in Unconstrained Environments,” in *Beijing University of Posts and Telecommunications, Tech. Rep 5.7 (2018)*, 2018.

[36] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia and S. Zafeiriou, “AgeDB: the first manually collected, in-the-wild age database,” in *IEEE conference on computer vision and pattern recognition workshops (CVPRW). IEEE.*, 2017.

[37] T. Zheng, W. Deng and J. Hu, “Cross-Age LFW: A Database for Studying Cross-Age Face Recognition in Unconstrained Environments,” in *Beijing University of Posts and Telecommunications Beijing 100876,China*, 2017.

[38] C. Whitelam , E. Taborsky, A. Blanton, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, J. Cheney and P. Grother, “IARPA Janus Benchmark-B Face Dataset,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE.*, 2017.

[39] B. Maze, J. Adams, . J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel , J. Anderson, J. Cheney and P. Grother, “IARPA Janus Benchmark – C: Face Dataset and Protocol,” in *international conference on biometrics (ICB). IEEE.*, 2018.

[40] H. Wu, S. Tian, A. Bhatta, J. Gutierre, G. Argueta, K. Ricanek Jr., . M. C. King and K. W. Bowyer, “Goldilocks Test Sets for Face Verification,” in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, 2026.

[41] N. Thom, A. DeBolt, L. Brown and E. Hand, “DoppelVer: A Benchmark for Face Verification.,” in *International Symposium on Advances in Visual Computing (ISVC), Lake Tahoe, NV, USA, Oct 2023.*, 2023.

[42] R. Zhang, P. Isola, A. A. Efros, E. Shechtman and O. Wang, “The Unreasonable Effectiveness of Deep Features as a Perceptual Metric,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[43] H. Wu and K. W. Bowyer, “Vec2Face+ for Face Dataset Generation,” 2025.

[44] H. Wu, J. Singh, S. Tian, L. Zheng and K. W. Bowyer, “Vec2Face: Scaling Face Dataset Generation with Loosely Constrained Vectors,” 2025.

[45] H. O. Shahreza and S. Marcel, “HyperFace: Generating Synthetic Face Recognition Datasets by Exploring Face Embedding Hypersphere,” 2024.

[46] U. D. o. H. Security, “Remote Identity Validation Rally,” 2024. [Online]. Available: https://www.dhs.gov/science-and-technology/remote-identity-validation-rally.

Draft

[47] S. Schellenberger, K. Shi, T. Steigleder, A. Malessa, F. Michler, L. Hameyer, N. Neumann, F. Lurz, C. Ostgathe, R. Weigel and A. Koelpin, “A Dataset of Clinically Recorded Radar Vital Signs with Synchronised Reference Sensor Signals,” *Scientific Data,* vol. 7, no. 291, 2020.

[48] F. P. Papantoniou, A. Lattas, S. Moschoglou, J. Deng, B. Kainz and S. Zafeiriou, “Arc2Face: A Foundation Model for ID-Consistent Human Faces,” in *European Conference on Computer Vision*, 2024.

[49] Q. Meng, S. Zhao, Z. Huang and F. Zhou, “MagFace: A Universal Representation for Face Recognition and Quality Assessment,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[50] P. Melzi, C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, D. Lawatsch, F. Domin and M. Schaubert, “GANDiffFace: Controllable Generation of Synthetic Datasets for Face Recognition with Realistic Variations,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2023.

[51] P. Grother, M. Ngan, K. Hanaoka, J. C. Yang and A. Hom, “Face Recognition Technology Evaluation (FRTE),” 2025.

[52] D. Geissbuhler, H. Otroshi Shahreza and S. Marcel, “Synthetic Face Datasets Generation via Latent Space Exploration from Brownian Identity Diffusion,” in *Proceedings of the International Conference on Machine Learning*, 2025.

[53] A. Desai, C. Freeman, Z. Wang and I. Beaver, “TimeVAE: A Variational Auto-Encoder for Multivariate Time Series Generation,” 2021.

[54] F. Boutros, J. H. Grebe, A. Kuijper and N. Damer, “IDiff-Face: Synthetic-Based Face Recognition through Fuzzy Identity-Conditioned Diffusion Model,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[55] Z. Wang, I. Samsten, I. Miliou, R. Mochaourab and P. Papapetrou, “guided locally constrained counterfactual explanations for time series classification,” *Machine Learning,* pp. Volume 113, 4639–4669, 2024.

[56] J. Černevičienė and A. Kabašinskas, "Explainable artificial intelligence (XAI) in finance: a systematic literature review," *Artificial Intelligence Review,* vol. 57, no. 216, p. 8, 2024.

[57] J. Sanz-Cruzado, N. Droukas and R. McC, “FAR-Trans: An Investment Dataset for Financial Asset Recommendation,” in *IJCAI*, 2024.

Draft

© 2024-2027 CERTAIN