11th International Conference of the Immersive Learning Research Network
Practitioner Proceedings iLRN2025

DOI: https://doi.org/10.56198/x5r5qk52

Ethics in Immersion: XR and the Synthetic Data Dance Revolution

Genevieve Smith-Nunes and Alex Shaw

University of Roehampton, London, UK
Glastonbridge Software, Edinburgh, Scotland
ges52@cantab.ac.uk

Abstract. This work in progress (WiP) paper is exploring immersive learning approaches within computing education, focusing on synthetic data (SD) generation, data ethics, biometrics for extended reality (XR). Using Google Colab™ with SynthDataVault this platform serves as both a practical tool for SD creation and an interactive teaching environment. With AR Ballet biometric datasets, students engage with SD generation concepts while critically examining ethical implications surrounding body-related data. The core educational objectives include developing AI competencies, applying SD within XR, fostering ethical awareness, and promoting interdisciplinary learning between computing and the arts. This approach integrates theoretical understanding with hands-on experience, addressing challenges such as differential privacy and algorithmic bias. Future stages will refine SD processes, enhance XR applications, and further explore the ethical dimensions of biometric data synthesis. By integrating technical and ethical considerations, the project aims to enhance computing education through cross-disciplinary, real-world creative and technical applications.

Keywords: XR, AR, Synthetic Data, Motion Capture, Computing Education.

1 Introduction

This work-in-progress (WiP) paper investigates the use of immersive learning approaches in computing education, particularly within synthetic data(SD) generation, data ethics, biometrics, and XR. Google Colab™, integrated with the SynthDataVault *library* [1], has to the potential to be an effective computing educational platform for generating synthetic data. The Colab™ platform supports two key functions: (i) providing a practical environment for SD creation and serving as an interactive teaching tool. (ii) The use of Python was prioritised due to its prevalence in pre-university computing education in England [2].

1.1 What is Synthetic Data

Synthetic data, first introduced by Rubin, is defined as 'data sets consisting of records of individual synthetic units rather than actual units' [3]. The European Union further describes synthetic data as "artificial data generated from original data and a model trained to reproduce the characteristics and structure of the original data," highlighting its role in secure data sharing and research.

The Agencia Española de Protección de Datos (AEPD) in line with EU directives, further defines synthetic data as "*artificial data that is generated from original data and a model that is trained to reproduce the characteristics and structure of the original data*" [4], highlighting its importance in secure data sharing and research.

1.2 AI Literacy in Contemporary Computing Education

AI literacy encompasses essential competencies enabling individuals to understand, critically engage with, and effectively use artificial intelligence, particularly within immersive technologies and learning environments. These competencies include recognising AI systems, understanding their capabilities and limitations, and using them ethically and responsibly. In immersive learning contexts, AI literacy involves understanding how AI

©2025 Immersive Learning Research Network

integrates with technologies such as extended reality (XR) including augmented reality (AR), virtual reality (VR). To create interactive, data-driven experiences that comply with curricular and exam requirements. AI literacy combined with immersive learning is a new area for pre-university computing education. Discovering the barriers and benefits is one of the drivers of this study.

2 Research Objectives and Problem Statement

Our WiP aims to address the growing need for practical, ethical AI and XR education in pre-university computing opportunities. The primary research question is: How can pre-university computing education effectively incorporate synthetic data generation tools to prepare students for emerging XR technologies?

2.1 Educational Objectives

Our work aims to addresses four key educational objectives:

1. Developing AI Competencies: Enhancing students' proficiency in AI and synthetic data generation within computing and XR contexts.
2. Synthesising and Using SD for XR: Enabling learners to generate and implement synthetic data in real-world XR applications.
3. Critical Thinking on Ethical AI: Encouraging ethical considerations around AI, including privacy, algorithmic bias, and responsible data use.
4. Interdisciplinary Learning: Combining computer science, arts, and immersive technologies.

These objectives directly respond to the challenges identified in contemporary computing education research [2] particularly regarding the integration of ethical considerations with technical skill development.

In support of these objectives, the development of the curriculum, in collaboration with participants, introduces learners and instructors to SD generation via a cross-disciplinary methodology that integrates creative computer science practice and theory, exposing students to real-world programming tasks and XR application development. This immersive pedagogical model balances theoretical understanding with hands-on experience, enabling learners to engage directly with key concepts in data ethics, such as differential privacy [4] while they observe the immediate effects of their work.

Developing AI competencies focuses on strengthening students' proficiency in AI and SD generation within computing and XR contexts. Synthesising and using SD for XR enables learners to create and apply SD in real-world XR applications. The SDs for 3d Modelling with tools such a Blender, or generative visual (e.g. particles effects), or audio effects. Critical thinking on ethical AI encourages ethical reflection on AI-related concerns, including privacy, algorithmic bias, and responsible data use. Interdisciplinary learning through arts and computing bridges computing and the arts through SD generation, particularly (in this case) via ballet and immersive technologies.

3 Background and Context

Our project uses biometric data from an in-development AR production, providing a unique computing educational resource that emphasises SD generation concepts. This immersive augmented reality ballet integrates synthetic data, biometrics, and AR to create data-driven learning opportunities. The narrative follows four astronauts on humanity's first journey beyond our solar system, explored through personal diaries, biometric sensors, and digital self-representations.

The deliberate use of synthetic data in one episode developed from collaborative design practuces involving authors, dancers, and choreographers, examining how immersive technology affects identity and human connection [4]. Through these elements, the AR ballet explores the evolving landscape of data, revealing both empowering and unsettling dimensions of digital augmentation in our lives. The educational learning opportunities arose from discussions and research of the AR in-production.

4 Methodology

The study follows a design-based research [5] and participatory design approach [6]. The technical methodology aligns with Agile development practices, integrating iterative feedback cycles [7]. Surveys and semi-structured

interviews inform the analysis, using interpretative phenomenological analysis and natural language processing [8].

4.1 Tool Selection and Justification

The development process begins with establishing the Colab™ environment for synthetic data generation. Following our established pipeline, raw data undergoes conversion to appropriate formats before being loaded into Colab™, with careful attention to metadata and primary key configuration. We selected the Gaussian Copula synthesiser model for this work, followed by a series of diagnostic tests that statistically compare the synthetic data with the original data. The final stage integrates with standard immersive experience development processes.The selection of Google Colab™ integrated with SynthDataVault [1] was based on several empirically supported factors:

1. Accessibility: Cloud-based platform requiring minimal setup, supported by research showing reduced technical barriers increase student engagement [9].
2. Python Integration: Chosen as the primary language due to its prevalence in UK pre-university computing education [2].
3. Interactive Learning: Notebook-style environment facilitating immediate feedback and experimentation.

4.2 Technical Pipeline

The specific process for generating synthetic data is illustrated in fig 1. This pipeline created the original human biometric (movement and EEG) datasets. These datasets were aggregated to provide more privacy for the individuals who provided their biometric data.

The SD generation process follows a structured pipeline. First, motion capture (BVH format) and electroencephalography (EEG, CSV format) data are gathered from ballet dancers performing choreographed movements. The raw biometric data is then cleaned, anonymised, and converted into standardised formats suitable for SD generation. This is currently being developed and tested over the next 5-7 months. The current design of the technical pipeline, Fig 1.

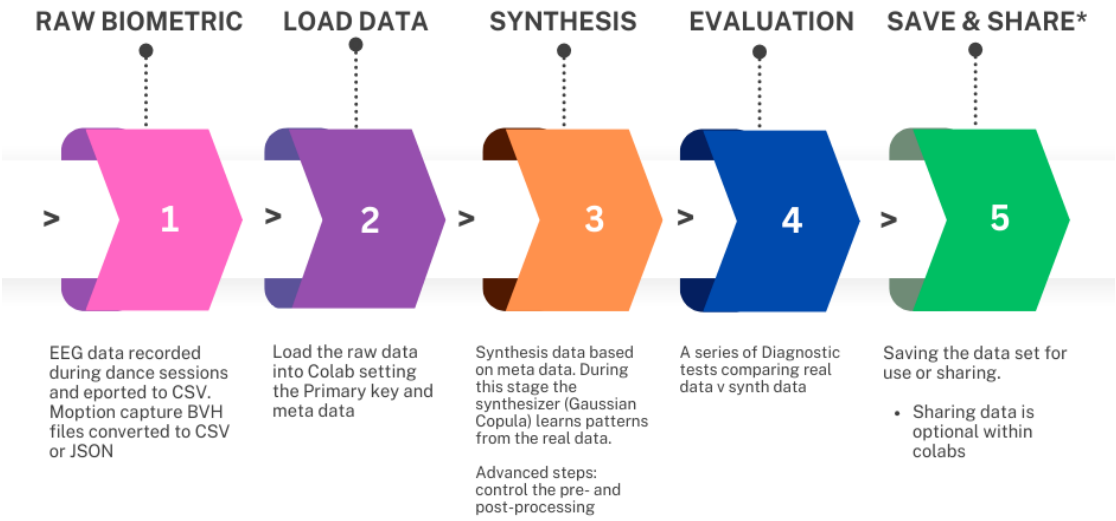

Fig. 1. Synthetics Data Generation Pipeline [10].

Firstly, capture raw biometric data, step 1 in Fig 1, in this study it is EEG data and Movement data which uses a multistep process. (i) video record movement dance at, least 60fps. (ii) Isolate dancer within video. (iii) This is the most technical stage: Using MoCapNET [11] which MocapNET uses two inference stages, the first one identifies 2D joint positions from an image, (frame from a video) and the second estimates the 3D pose of the human from 2D joint positions. This final inference result can be exported as BVH file format and applied to rigged 3D character models in tools such as Blender™ or other 3D modelling software.

Next, Gaussian Copula-based generative models are chosen due to their capability to maintain statistical correlations within the data. By using SynthDataVault, SD the aim is to replicate the statistical properties of the

original dataset while ensuring privacy. The generated SD is then statistically compared with real data using similarity metrics, such as Kullback-Leibler divergence and mean absolute error [1]. Finally, the validated synthetic dataset will be integrated into immersive environments. The two proposed routes are 3D model generation, in Blender™, from SD and the use of SD for effects within an XR environment.

Immersive development tool selection from SD will be participant lead due to variations of access for all participants. We aim to use opensource (at the point of use) where possible and limit the installation of software due to site network barriers for adding new or additional software. We plan to start with brainwave EEG SDs for effects within the XR platforms selected by the participants. From this we aim to develop specific educational resources for both EEG and movement SDs integration. We believe that SD movement will not function as well, if at all, as capture human movement datasets.

5 Study Structure

The development process begins with establishing the Colab™ environment for synthetic data generation. Following our established pipeline, raw data undergoes conversion to appropriate formats before being loaded into Colab™, with careful attention to metadata and primary key configuration. Following the testing with participants on generating SDs they will then work to develop 3D models or use the datasets with 3D immersive environments for additional effects. The tools will be participant selected and appropriate for their course or educational keystage.

- December – April: Development of Colab™ environment
- April-July: Testing workflow with pre-service computing teachers and undergraduate students. Generation of SDs followed by XR 3D model development testing from SD.
- Following analysis on fieldwork, commencement of implementation of validation metrics for educational impact.
- Refinement of synthetic data generation processes, practices, and documentation for teaching and learning use, tailored to England pre-university contexts.

6 Next Steps and Future Work

As we at the early stage – design phase at the time of writing the upcoming phases of our project will continue refining the synthetic data design process for creative computing applications. From April to July, we will conduct extensive testing of our workflow for creating synthetic datasets, working with pre-service secondary computing teachers and a select group of computer science undergraduates. This testing phase will focus on validating our approaches and refining our educational resources.

Beginning in July, we will undertake a comprehensive examination of the ethical implications surrounding the collection, synthesis, and use of body-related data in XR applications. This research aims to define the limitations of synthetic data generation for immersive learning within the context of computing education. Develop educational resources that effectively articulate both technical processes and ethical considerations.

The ethical considerations surrounding synthetic data synthesis in this project aim not only address the technical and legal challenges of working with sensitive information but also foster a reflective, ethical, and responsible approach to AI-enhanced immersive learning environments. This plural focus aims to ensure that both educators and students are equipped to make informed decisions in their future roles as creators, users, and evaluators of emerging technologies.

References

1. Patki, N., Wedge, R., Veeramachaneni, K.: The Synthetic Data Vault. In: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), pp. 399–410. IEEE, Montreal, QC, Canada (2016).
2. Hadwen-Bennett, A., Kemp, P.: Programming in Secondary Education in England: Technical Report. King's College London (2024).
3. Rubin, D.: Discussion: Statistical disclosure limitation. J. Off. Stat. 9, 461–468 (1993).
4. EDPS: Synthetic data and data protection | AEPD, https://www.aepd.es/en/prensa-y-comunicacion/blog/synthetic-data-and-data-protection, last accessed 2023.
5. Bakker, A.: Design Research in Education: A Practical Guide for Early Career Researchers. Routledge (2019).
6. Bayley, A.: Posthuman Pedagogies in Practice: Arts-Based Approaches for Developing Participatory Futures. Springer (2018).

7. Knaster, R., Leffingwell, D.: SAFe 5.0 Distilled: Achieving Business Agility with the Scaled Agile Framework. Addison-Wesley Professional (2020).
8. Smith, J., Flowers, P., Larkin, M.: Interpretative Phenomenological Analysis. Sage Publications Ltd (2021).
9. Gottschalk, F., Weise, C.: Digital equity and inclusion in education: An overview of practice and policy in OECD countries (2023).
10. Smith-Nunes, G., Shaw, A.: Dancing with Synthetic Data: AI Educational Research using an AR Ballet. In: Proceedings of the 30th UK Academy for Information Systems (UKAIS) International Conference (2025).
11. Qammaz, A., Argyros, A.: MocapNET: Ensemble of SNN Encoders for 3D Human Pose Estimation in RGB Images, https://users.ics.forth.gr/~argyros/mypapers/2019_09_BMVC_mocapnet.pdf, last accessed 2019.