Synthetic fields, real gains

Enhancing smart agriculture through hybrid datasets

Paul Wachter [1], Niklas Kruse [2] and Julius Schöning [2]

Abstract: Artificial intelligence (AI) promises transformative impacts on society, industry, and agriculture, while being heavily reliant on diverse, quality data. The resource-intensive "data problem" has initialized a shift to synthetic data. One downside of synthetic data is known as the "reality gap", a lack of realism. Hybrid data, combining synthetic and real data, addresses this. The paper examines terminological inconsistencies and proposes a unified taxonomy for real, synthetic, augmented, and hybrid data. It aims to enhance AI training datasets in smart agriculture, addressing the challenges in the agricultural data landscape. Utilizing hybrid data in AI models offers improved prediction performance and adaptability.

Keywords: hybrid data, synthetic data, augmented data, smart farming, reality gap

1 Introduction

Artificial intelligence (AI) has been one of the most prominent topics in recent years. It has emerged as an essential technology in agriculture, enabling advancements in crop management [SWT23], soil analysis [Wa23], yield prediction, and identification of plant diseases and weeds, among others. One fundamental aspect of this technology is the importance of the data used to train these systems. Acquiring, creating, preparing, and labeling this data can incur significant costs, known as the "data problem" [Ni21]. For example, data sparsity, i.e., missing data, or class imbalance, i.e., a skewed distribution of data, are resulting problems that are highly common in the agricultural sector [Co22]. One solution to this problem is synthetic data. The IT research firm and consultancy Gartner predicts that in 2030, 60 % of all data used to train AI systems will be synthetically generated [GA23]. While cost-effective, synthetic data lacks realism, termed the "reality gap" [Ni21]. Hybrid data addresses this by combining synthetic and real data by leveraging cheap synthetic data and bridging the gap with a smaller amount

[1] Osnabrück University, Institute of Cognitive Science, Neuer Graben 29, 49074, pwachter@uni-osnabrueck.de, https://orcid.org/0000-0002-6224-6140

[2] Osnabrück University of Applied Sciences, Faculty of Engineering and Computer Science, Albrechtstr. 30, 49076 Osnabrück, niklas.kruse@hs-osnabrueck.de, https://orcid.org/0009-0001-7080-6662; j.schoening@hs-osnabrueck.de https://www.orcid.org/0000-0003-4921-5179

of real data [Fa19]. However, inconsistent terminology in current literature remains a challenge.

2 Existing taxonomy of real, synthetic, augmented, and hybrid

In the following literature review, the terminology for different data types is inconsistently used, except for “real data”. This section clarifies the definitions of “real data”, “synthetic data”, “augmented data”, and “hybrid data” to address these inconsistencies.

2.1 Real data

The Cambridge Academic Content Dictionary defines data as “information collected for use” [CA08], underscoring the essential aspect of real data obtained from the physical world through observations or measurements. Real data originates from real-world sources, including sensors, devices, and human interactions. Importantly, this definition is consistently used in all literature sources we have examined.

2.2 Augmented data and synthetic data

The fundamental distinction in the definitions of augmented data lies in whether it includes or excludes synthetic data. An example of the first is given by [Ni21], where augmented data is defined as “[...] transformations of the input data that change the target labels in predictable ways”. In contrast, synthetic data, as per [Ni21], is entirely generated manually or by an algorithm, which may include generation through generative AI. Despite the training of these algorithms on real data, their output is considered synthetic, not augmented data. In this context, augmented data consists of transformed real data, while synthetic data comprises entirely generated data.

The second interpretation, where augmented data includes synthetic data, can be found in [GCB23]: “[augmented data is] data that has been altered to include extra information. Synthetic data is data that has been created artificially, using data augmentation techniques”. Here, synthetic data is explicitly categorized as a subset of augmented data.

2.3 Partially synthetic and hybrid data

Partially synthetic and hybrid data refer both to a fusion of real and synthetic data. Again, differing interpretations exist. [EMH20] defines partial synthetic data as a mix of artificially generated and retained parts of real data, resembling traditional de-identification. [EMH20] characterizes hybrid data similarly, blending synthetic and real data, but with the synthetic part based on theoretical understanding, even if actual data is lacking. Thus, partially synthetic data masks real data, while hybrid data introduces new synthetic elements. In practice, [EMH20]'s definition of partially synthetic data is widely used, particularly in disclosure control for tabular and sensitive data [Jo22; GLR22]. Its application to other data types, like images, is less common, cf. [GZ18].

The prevalent definition of hybrid datasets, per [Ni21], is "using synthetic data to augment existing real datasets". Unlike [EMH20], [Ni21] specifies that the term hybrid is applicable to datasets, not individual data samples. Lastly, [TO23] uses the term hybrid to describe datasets, but here, hybrid datasets contain only synthetic samples, which are selected based on similarity to samples in a real dataset.

3 Consistent taxonomy of real, synthetic, augmented, and hybrid

Ensuring effective knowledge transfer from academia to practical applications in the agricultural sector requires consistent terminology. Therefore, a unified terminology and taxonomy for categorizing data types is proposed in Figure 1.

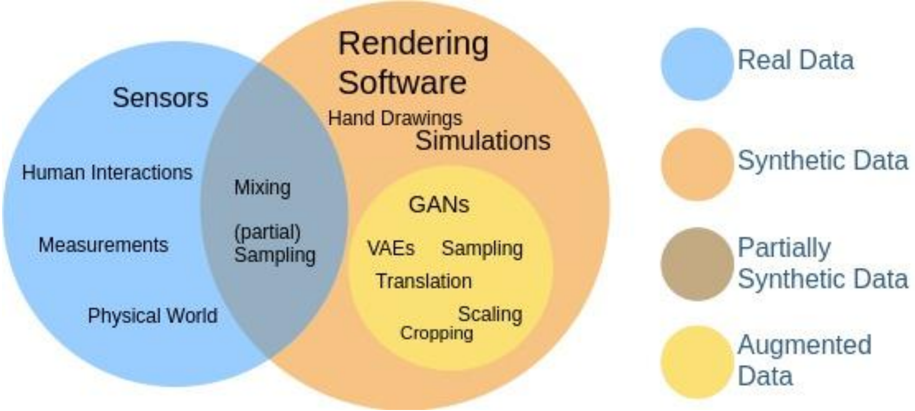

Fig. 1: Taxonomy of real, partially synthetic, synthetic, and augmented data

We define real data as information collected from the physical world, as presented in Section 2.1, while synthetic data is all artificially generated and transformed data. This includes manual processes, mathematical models, algorithms, and simulations. We define augmented data as a subset of synthetic data. Augmented data samples are transformations of real data samples, which can be basic modifications, like cropping and rotation, as well as more complex transformations, such as those achieved through generative AI. This means we define the data samples produced by generative AI as transformations of the data used to train the generative AI, i.e., augmented data. If a dataset contains both real and augmented samples, it is an augmented dataset. We define partially synthetic only for data samples. They possess real and synthetic attributes; one part originates from real sources, while the other is synthetically generated. Lastly, we define the term hybrid only for datasets, not samples. A hybrid dataset contains both real and synthetic samples; the latter are not augmented data samples. This is a crucial distinction because it would be considered an augmented dataset without this condition. It is important to note that if it meets this criterion, a dataset can still include augmented data samples while being labeled as hybrid data.

4 Benefits of hybrid data

The main challenge with synthetic data is the "reality gap" [Ni21], also known as the "domain gap" [Wa19], stemming from differences from real data. This gap causes a significant drop in prediction performance when models trained on synthetic data are deployed in real-world scenarios. Domain adaption techniques try to adapt the synthetic to the real to close this gap. One domain adaption technique is hybrid data, which utilizes cheap and in large quantities available synthetic data while containing real data to bridge the gap to the actual environment. On the one hand, this improves the model's prediction performance, and on the other, it reduces the need for real data.

4.1 Required amount of data and model's prediction performance

Using hybrid data is particularly valuable in scenarios that suffer from the data sparsity problem, where producing a sufficient volume of real data is impractical. Various publications propose and assess a data ratio of 10 % real data and 90 % synthetic data. This ratio has demonstrated notable significance in enhancing model prediction performance, often outperforming models trained only on multiple times the amount of real data. For instance, in [RHB17], this 10 % real and 90 % synthetic data ratio was utilized for their best-performing model. In [Ge17], using only 10 % of the available real data, combined with 100 % of the synthetic data, yielded higher or comparable detection accuracies than when trained on 100 % of the real data and 0 % of the synthetic data. A similar result is obtained in [Fa19]. In general, as the quantity of real data increases, there is an improvement in model prediction performance. However, even in scenarios where a large amount of real data is available, introducing hybrid data continues to enhance model prediction performance, as demonstrated in [Tr18] and [Fa19]. Lastly, a study conducted by [Wa19] examines the effectiveness of hybrid data in contrast to another domain adaptation method, which was based on a Structural Similarity Index embedding cycle GAN, and shows that the former yielded superior prediction performance.

4.2 Use in agriculture

Data sparsity and class imbalance are frequently encountered in agricultural data, cf. [Co22], [SWT23]. Considering the results presented above, the problem of data sparsity can be mitigated by combining the existing real datasets with synthetic ones. An example of this is given by [WMH18]. Here, a leave segmentation model was trained on hybrid data, consisting of only 128 real images combined with 10.000 synthetic images. The resulting model outperformed all the state-of-the-art approaches.

Similarly, in [CPN20], a model for plant-weed segmentation trained using real, purely synthetic, and hybrid data was assessed. The number of used samples was 734 real samples, 734 synthetic samples, and 734 synthetic plus 100 real samples. While the model trained only on synthetic samples achieved significantly lower prediction performance compared to the model trained on only real samples, the model trained on the hybrid dataset achieved comparable prediction performance. However, only a tiny amount of the real data was added to the synthetic data. Lastly, [GCK20] showed that a hybrid data approach can be used in scenarios where the data exhibits class imbalance. On the task of strawberry detection, they improved their models F1-scores, which are typically used to quantify the prediction performance in cases of class imbalance, using hybrid data. The model's overall accuracy was neither improved nor harmed, but the model's bias towards one class could be reduced.

5 Conclusion

The "data problem" is one of the fundamental problems in deploying AI systems in agriculture. Synthetic data is a promising and cost-effective alternative, but it comes with the problem of the "reality gap". Hybrid data, a combination of synthetic and real data, has shown to be a promising solution to bridge the reality gap and enhance AI model prediction performance and adaptability. AI models can achieve impressive results by using just a small amount of real data combined with synthetic data, sometimes even outperforming those trained on a greater quantity of real data. In the context of agriculture, where problems like data sparsity or class imbalance are widespread, it is possible to significantly improve the prediction performance of AI models and reduce their biases with the use of hybrid data, making them more adaptable to real-world scenarios. A unified taxonomy of the different data types clarifies their differences and similarities, ensuring their potential is fully realized in agricultural production. As next steps, we will use this taxonomy to create hybrid data sets and research methods for using hybrid data for AI-based agrarian applications.

Bibliography

[CA08] Cambridge Academic Dictionary. Cambridge University Press, Cambridge, 2008.

[Co22] Condran, S. et. al.: Machine Learning in Precision Agriculture: A Survey on Trends, Applications and Evaluations Over Two Decades. IEEE Access, vol. 10, S. 73786-73803, 2022.

[CPN20] Carbone, C.; Potena, C.; Nardi, D.: Simulation of near Infrared Sensor in Unity for Plantweed Segmentation Classification. In: International Conference on Simulation and Modeling Methodologies, Technologies and Applications - Volume 1: SIMULTECH, S.81-90, 2020.

[EMH20] El Emam, K.; Mosquera, L.; Hoptroff, R.: Practical Synthetic Data Generation. O'Reilly Media, Inc, Sebastopol, 2020.

[Fa19] Farzan, E.N. et. al.: How much real data do we actually need: Analyzing object detection performance using synthetic and real data. CoRR., 2019.

[GA23] Gartner, https://blogs.gartner.com/andrew_white/2021/07/24/by-2024-60-of-the-data-used-for-the-development-of-ai-and-analytics-projects-will-be-synthetically-generated/, Accessed: 15.12.2023.

[GCB23] Gürsakal, N.; Çelik, S.; Birişçi E.: Synthetic Data for Deep Learning: Generate Synthetic Data for Decision Making and Applications with Python and R. Apress, Berkeley CA, 2023.

[GCK20] Goondram, S.; Cosgun, A.; Kulić, D.: Strawberry Detection using Mixed Training on Simulated and Real Data. ArXiv, abs/2008.10236, 2020.

[Ge17] Georgakis, G. et al.: Synthesizing training data for object detection in indoor scenes. ArXiv, 2017.

[GLR22] Grund S.; Lüdtke O.; Robitzsch A.: Using synthetic data to improve the reproducibility of statistical results in psychological research. Psychological Methods, 2022.

[Jo22] Jordon, J. et. al.: Synthetic Data - what, why and how? ArXiv, 2022.

[Ni21] Nikolenko, S.I.: Synthetic Data for Deep Learning. Springer Cham, 2021.

[RHB17] Rajpura, P.S.; Hegde. R.S.; Bojinov, H.: Object Detection Using Deep CNNs Trained on Synthetic Images. ArXiv, 2017.

[SWT23] Schöning, J.; Wachter, P; Trautz, D.: Crop rotation and management tools for every farmer?: The current status on crop rotation and management tools for enabling sustainable agriculture worldwide. Smart Agricultural Technology 3, S. 100086, 2023.

[TO23] Towards Data Science, https://towardsdatascience.com/synthetic-data-key-benefits-types-generation-methods-and-challenges-11b0ad304b55, Accessed: 15.12.2023.

[Tr18] Tremblay, J. et al.: Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops (CVPRW), 2018.

[Wa19] Q. Wang, et. al.: Learning From Synthetic Data for Crowd Counting in the Wild. In Conference on Computer Vision and Pattern Recognition (CVPR), S. 8190-8199. Long Beach, CA, USA, 2019.

[Wa23] Wachter, P. et al.: A Smartphone App for Simple Soil Structure Analysis. IEEE International Humanitarian Technology Conference (IHTC) 2022.

[WMH18] Ward, D.; Moghadam, P.; Hudson, N.: In: Proceedings of the British Machine Vision Conference (BMVC) Workshop on Computer Vision Problems in Plant Pheonotyping (CVPPP), 2018.