# Introduction

With the advent of data collection methods and the meteroric rise of data science in part due the development of better system architectures, machine learning has come to play an important part in manufacturing. Global data creation has increased exponentially in the past few years. Global DataSphere forecast[[1]](#footnote-1), 2021-2025, predicts that Data gathering will increase by 23% annually over the forecast period. This is adding to the tenfold increase in data creation over the past decade. The manufacturing industry has had a significant part in this with industry 4.0, internet of things helping the push towards constant improvement. In this study, the possiblity of using machine learninf to improve the performance of a hydraulic press is studied.

In this study a hydraulic press with a cylindrical drive is considered. (*details to be included later*). Our goal here is to optimize the condition monitoring capabilities. Condition monitoring, is a approach to maintenance which uses available sensor data to deduce the health of the systems. It focuses on early detection of abnormal behaviour and works to eliminate failures in the operation of the system. Catastrophic failures and the cost involved in fixing it can be reduced or eliminated using better condition monitoring systems.

(Make sure that the toolbar ‚IFAS’ is activated. If the toolbar does not occur then correct the security level (‚Extras -> Makro -> Sicherheit’) and open this template again.)

## Data

The most important feature of any machine learning mode is the quality and the quantity of the data available. There are considerable costs involved in the generation and collection of read world data, along with the practical inability to gather data under certain conditions. Because of this there might exist a class imbalance in the available data set. Class imbalance is the over representation of certain classes and the under representation of certain classes. E.g., it would be costly to gather failure data of a manufacturing system as whole or a component of it, multiple times in the real world. Hence in most real-world datasets there might be a mismatch in the classes. To augment the available data or in some cases serve as a stand in for real-world data synthetic data is used. Synthetic data is generated artificially by a simulation model that is designed to replicate the operation of the real-world system. An excellent example of the use of synthetic data can be understood from the work of creating driving data using GTA V game engine for use in connected driving models.[[2]](#footnote-2) However synthetic data is not absolutely similar to real-world data. The difference between real world data and synthetic data is called the reality-gap. The main concern of this study is to research the different ways to quantify and close the reality gap between the synthetic data generated using a 1-D simulation of the Hydraulic press architecture in Modelica-based simulation tool SimulationX and the real-world data available through experimental data gathering.

## Possible extrapolations

Some possible areas to be explored:

### Transfer Learning

Transfer learning is the process by which a model trained using a selected dataset can be used to resolve a different dataset that is related to the one used to train the model. Particular area of interest is Domain adaptation where the model deals with domain shift. When the data distribution of the training dataset and the distribution of the testing or deployed data are different this results in a domain shift. Domain adaption deals with datasets having same feature space unlike transfer learning which can be used on datasets with differing feature spaces.

### Data Augmentation methods

Some data augmentation methods can be used to correct the class imbalance, such as Random under sampling, over sampling, SMOTE (Synthetic Minority Oversampling TEchnique), etc. usefulness and effectiveness used along with previous methods need to be researched further.

(Use the template „level 4“ to insert a fourth rubric level.)

(To insert figures put the cursor to the right position. Then click on the smiley (‚past figure’) and select the suited file. Click ‚Ok’ and insert the figure caption.)

# State of the Art

Asasas

## Previous work done using the available databases

The data available from the Hydraulic system is in the form of a multivariate, cyclic, non-stationary time series data. The following is a summary of the work previously done using the same or related database. Synthetic data is generated along the lines of real-world data, using Simulation X which is a Modelica based simulation tool. A 1D simulation of the hydraulic circuit used in the press is modelled and is used in data generation[[3]](#footnote-3). Faults such as cylinder leakage, friction, worn edges are modelled by varying parameters of the components to recreate the effects of the faults in real-world data. In the next study a mixed domain database containing both real-world data and synthetic data is used in the machine learning model.[[4]](#footnote-4) The accuracy of the model is evaluated using six different ratios according to which the real-world data and the synthetic data are mixed. An increase in the synthetic data added to the real -world data provides more coverage and instances of a particular class but at the same time due to the reality gap, there exists a difference between the two data which induces noise in the resulting mixed domain database. The times series data from manufacturing systems usually have unequal lengths and as such will prove difficult to be used in machine learning models, hence feature extraction needs to be done. Using *tsfresh*, 16 base features are generated.[[5]](#footnote-5) Feature engineering is applied to the generated features adding local information and trend information and selective use of features directly affecting the model. Another method of reducing the high dimensional time series data is performed using variational autoencoders (VAE)[[6]](#footnote-6). The VAE works to compress the data achieving a 99.94% compression rate, while a LSTM (Long Short-Term Memory) network is used to preserve the time component of the data.

## Metrics for differentiation of data

The first step in solving the problem of a reality gap is to quantify the gap between the two datasets. This step is easier for image data compared to tabular data. Synthetic data can be generated from different sources but evaluating tabula or sequential data is challenging as even two random selection of the same dataset, even if it is real world might not be the same or have the same distribution. One study classifies the utility metrics used to measure the data, into two main categories. Broad metrics are used to quantify the database, the distribution similarities, etc., while the narrow metrics measure the ability of the synthetics data generators to generate data that can be used to replicate certain analysis. F. K. Dankar et al[[7]](#footnote-7) further introduce further categories to broad metrics based on the statistics they compute namely, attribute fidelity, bivariate fidelity and population fidelity. Attribute fidelity measures the basis structural similarity between the two datasets, such as variable types, formats, names, means and ranges. Bivariate fidelity compares the relationship between different variable with the same from the real-world database. This is necessary in cases where the relationship between different variable affect the model. For example, the relationship between age and heart failure should be replicated in a synthetic database trying to fill in for the real-world database. Population fidelity compares the whole databases. This provides a measure of how well the synthetic data compares to the real-world data in its entirety. Application Fidelity a type of narrow metric helps to evaluate how well a synthetic dataset perform in a specific analysis or application.

|  |  |  |
| --- | --- | --- |
| **Utility** | **Measurements** | **Example metrics** |
| Broad measure: Univariate fidelity | Structural similarity | variable types, formats, names, means and ranges |
| Univariate distribution | Hellinger, graphical methods, Normalized Kullback-Leibler divergence |
| Broad measure: Bivariate fidelity | Bi-variable correlation | Pairwise correlation difference, mutual and heat maps |
| Broad Measure: population fidelity | difference in statistical dependency structure (multi variate correlation) | cross classification |
| cluster based distributional difference. | Log-cluster |
| difference in empirical distributions | Kolmogorov-Smirnov type stations |
| Likelihood metrics | bayessian netwoek or Gaussian mixture models |
| Distinguishability | Propensity |
| Narrow Measure: Application fidelity | predication accuracy | ability to replicate studies performed on real data |

[[8]](#footnote-8)

Table 2.2 Data Utility Metrics

S Chundawat et al,[[9]](#footnote-9) provide a new metric TabSynDex, claimed to be reliable and can perform robust measurement of synthetic data similarity. TabSnDex combines five component scores to perform the analysis. Whereas, in a simplified version, El Emam et al, [[10]](#footnote-10) compare the performance of six different metrics to compare a synthetically generated health data to the real-world data. Of the following six, Maximum Mean Discrepancy, Multivariate Hellinger Distance, Wasserstein Distance, Cluster Analysis Measure, mean square error metric, Workload Aware (Narrow) Metrics, they conclude Hellinger distance is best suited to measure SDG (synthetic data generator)’s usefulness.

## Bridging the reality gap.

From the problem at hand, there is a need to transform the synthetic data to be more representative of the real-world or experimental data. A lot of work has already been done in this field, mostly in the fields of computer vision. But there exists a possibility that some if not most of the methods can be used on non-image data too. New research in the field of computer vision, connected driving, augmented reality etc requires huge labelled datasets, most of which are costly to obtain, label and come with legal and privacy concerns. So simulated data is used to replace the real-world data. But the domain-gap or reality-gap is a considerable obstacle that needs to be overcome. One of the methods used to overcome this is domain randomization (Valtchev and Wu 2021; Tremblay et al.),[[11]](#footnote-11) where the model is forced to learn across different domains so that it learns the deeper truth in the data without consideration of the domain. Generative Adversarial Networks (GANs) are playing a significant role in replicating or introducing style transfer in image data.[[12]](#footnote-12)[[13]](#footnote-13) GANs consist of two neural networks, a generator and a discriminator competing against each other. The generator generates image data out of random noise or pretrained conditions and a discriminator tries to discriminate the generated image and the real-world image. The loss function is back propagated. To be more specific conditional GANs[[14]](#footnote-14) are important in the fixing the reality-gap. Conditional GANs accept a conditional input into the generator on the generation of image or other forms of data.

### Tabular data

Tabular data is structured data that can be represented in rows and columns, though the training and use of tabular data in machine learning models is easier than image data, generation or style transfer gets complicated. Tabular data may be numerical (continuous or discrete) or categorical (ordinal or nominal). The use of GANs on tabular data is on the increase. GANs are used to generate tabular data[[15]](#footnote-15), conditional GANs[[16]](#footnote-16) can also been used. Continuous data can also be generated[[17]](#footnote-17), RelGAN can be used for text generation[[18]](#footnote-18). Time series data can also be generated or used in style transfer models. Da Silva et al[[19]](#footnote-19) use Denoising AutoEncode (DAE) after experimenting with GANs and VAE, to style transfer financial data. GANs are still experimental and training is prone to collapse.

El Laham, et al[[20]](#footnote-20) propose a style transfer for synthetic time series generation. Style transfer is a term used in image generation, each image has a style that can be represented by a Gram matrix of its features. These are the features that are learned by the different layers of a CNN network. In image data GANs are often used to transfer style from one image to another[[21]](#footnote-21), it is possible to recreate artistic styles by merging two different scenes. In this study a novel algorithm called StyleTime is proposed which can be used to learn the trend from one time series data and combine it with the style of another dataset.

## Condition Monitoring

Publication bibliography

Aziira, A. H.; Setiawan, N. A.; Soesanti, I. (2020): Generation of Synthetic Continuous Numerical Data Using Generative Adversarial Networks. In *J. Phys.: Conf. Ser.* 1577 (1), p. 12027. DOI: 10.1088/1742-6596/1577/1/012027.

Brumand-Poor, Faras; Makansi, Faried; Liao, Jiakun; Schmitz, Katharina (2015): Implementation of a Variational Autoencoder for Dimension Reduction of a Hydraulic System. In *AMM* 764-765, pp. 613–618. DOI: 10.4028/www.scientific.net/AMM.764-765.613.

Chundawat, Vikram S.; Tarun, Ayush K.; Mandal, Murari; Lahoti, Mukund; Narang, Pratik (2022): A Universal Metric for Robust Evaluation of Synthetic Tabular Data. In *IEEE Trans. Artif. Intell.*, pp. 1–11. DOI: 10.1109/TAI.2022.3229289.

Da Silva, Brandon; Shi, Sylvie Shang: Style Transfer with Time Series: Generating Synthetic Financial Data.

Dankar, Fida K.; Ibrahim, Mahmoud K.; Ismail, Leila (2022): A Multi-Dimensional Evaluation of Synthetic Data Generators. In *IEEE Access* 10, pp. 11147–11158. DOI: 10.1109/ACCESS.2022.3144765.

El Emam, Khaled; Mosquera, Lucy; Fang, Xi; El-Hussuna, Alaa: Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study.

El-Laham, Yousef; Vyetrenko, Svitlana: StyleTime: Style Transfer for Synthetic Time Series Generation, pp. 489–496. DOI: 10.1145/3533271.3561772.

Global DataSphere forecast.

GTA VP2 a Dataset for Multi-Agent Vehicle Trajectory Prediction.

Makansi, Faried: Feature Generation and Evaluation for Data-Based Condition Monitoring of a Hydraulic Press\_v4.

Makansi, Faried; Schmitz, Katharina (2021): Simulation-Based Data Sampling for Condition Monitoring of Fluid Power Drives. In *IOP Conf. Ser.: Mater. Sci. Eng.* 1097 (1), p. 12018. DOI: 10.1088/1757-899X/1097/1/012018.

Makansi, Faried; Schmitz, Katharina (2022): Fault Detection and Diagnosis for a Hydraulic Press by Use of a Mixed Domain Database. In : BATH/ASME 2022 Symposium on Fluid Power and Motion Control. BATH/ASME 2022 Symposium on Fluid Power and Motion Control. Bath, UK, 14/09/2022 - 16/09/2022: American Society of Mechanical Engineers.

Nie, Weili; Narodyska, Nina; B. Patel, Ankit: RelGAN: RELATIONAL GENERATIVE ADVERSARIAL NETWORKS FOR TEXT GENERATION.

Prakash, Aayush; Boochoon, Shaad; Brophy, Mark; Acuna, David; Cameracci, Eric; State, Gavriel et al.: Structured Domain Randomization: Bridging the Reality Gap by Context-Aware Synthetic Data.

Rojtberg, Pavel; Pollabauer, Thomas; Kuijper, Arjan: Style-transfer GANs for bridging the domain gap in synthetic pose estimator training, pp. 188–195. DOI: 10.1109/AIVR50618.2020.00039.

Sohaliya, Gaurav; Sharma, Kapil: An Evolution of Style Transfer from Artistic to Photorealistic: A Review, pp. 1–7. DOI: 10.1109/ASIANCON51346.2021.9544924.

Soni, Rajat Kumar; Nair, Binoy B.: Deep Learning Based Approach to Generate Realistic Data for ADAS Applications, pp. 1–5. DOI: 10.1109/ICCCSP52374.2021.9465529.

Tremblay, Jonathan; Prakash, Aayush; Acuna, David; Brophy, Mark; Jampani, Varun; Anil, Cem et al.: Training Deep Networks with Synthetic Data: Bridging the Reality Gap by Domain Randomization, pp. 1082–10828. DOI: 10.1109/CVPRW.2018.00143.

Valtchev, Svetozar Zarko; Wu, Jianhong (2021): Domain randomization for neural network classification. In *Journal of big data* 8 (1), p. 94. DOI: 10.1186/s40537-021-00455-5.

Wang, Ting-Chun; Liu, Ming-Yu; Zhu, Jun-Yan; Tao, Andrew; Kautz, Jan; Catanzaro, Bryan: High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs, pp. 8798–8807. DOI: 10.1109/CVPR.2018.00917.

Xu, Lei; Skoularidou, Maria; Cuesta-Infante, Alfredo; Veeramachaneni, Kalyan: Modeling Tabular data using Conditional GAN.

Xu, Lei; Veeramachaneni, Kalyan: Synthesizing Tabular Data using Generative Adversarial Networks.

1. Global DataSphere forecast. [↑](#footnote-ref-1)
2. GTA VP2 a Dataset for Multi-Agent Vehicle Trajectory Prediction. [↑](#footnote-ref-2)
3. Makansi and Schmitz 2021. [↑](#footnote-ref-3)
4. Makansi and Schmitz 2022. [↑](#footnote-ref-4)
5. Makansi. [↑](#footnote-ref-5)
6. Brumand-Poor et al. 2015. [↑](#footnote-ref-6)
7. Dankar et al. 2022. [↑](#footnote-ref-7)
8. Dankar et al. 2022. [↑](#footnote-ref-8)
9. Chundawat et al. 2022. [↑](#footnote-ref-9)
10. El Emam et al. [↑](#footnote-ref-10)
11. Prakash et al. [↑](#footnote-ref-11)
12. Rojtberg et al. [↑](#footnote-ref-12)
13. Soni and Nair. [↑](#footnote-ref-13)
14. Wang et al. [↑](#footnote-ref-14)
15. Xu et al. [↑](#footnote-ref-15)
16. Xu and Veeramachaneni. [↑](#footnote-ref-16)
17. Aziira et al. 2020. [↑](#footnote-ref-17)
18. Nie et al. [↑](#footnote-ref-18)
19. Da Silva and Shi. [↑](#footnote-ref-19)
20. El-Laham and Vyetrenko. [↑](#footnote-ref-20)
21. Sohaliya and Sharma. [↑](#footnote-ref-21)