AHFE International

Model Training Through Synthetic Data Generation: Investigating the Impact on Human Physical Fatigue

Arsalan Lambay[1], Phillip L Morgan[2], Ying Liu[3], and Ze Ji[4]

[1,3&4]School of Engineering, Cardiff University, Cardiff, Wales, UK, CF24 3AA

[2]School of Psychology, Cardiff University, Cardiff, Wales, UK, CF10 3AT

ABSTRACT

Collaborative robots, or cobots, are one of the Industry 4.0 technologies that have and continue to change many industrial procedures. However, amid this technological advancement, the persisting physical strain on human workers remains a significant concern. Even with the advent of cobots aimed at alleviating burdensome tasks, certain physical jobs continue to induce fatigue in human workers. Addressing this challenge necessitates the development of robust solutions that combine technological innovation with human-centric considerations. One critical aspect in mitigating physical fatigue in human workers involves the application of Machine Learning (ML) models. These models heavily depend on data obtained from real-world situations that accurately represent the complexities of physical strain. However, this kind of data is frequently limited and costly to gather using sensors, which hinders the development of an effective ML model. This scarcity underscores the need for alternative approaches, with Synthetic Data Generation (SDG) emerging as a viable solution to this problem. The production of synthetic data offers a new approach to address the lack of relevant data needed to train machine learning algorithms. By employing techniques like Tabular Generative Adversarial Networks (GANs), synthetic datasets can be created, simulating realistic human physical fatigue detection features. Tabular GANs have, for example, been shown to be effective in creating synthetic data that closely resembles the statistical characteristics and patterns of real-world datasets. Furthermore, tabular GANs present a scalable and affordable response to the problem of data scarcity. The research reported here presents a novel approach centred on employing the Tabular GAN methodology to create synthetic datasets encompassing key features pertinent to the detection of human physical fatigue. The results of this study are expected to contribute substantially to creating robust solutions to alleviate physical strain and enhance human workers' overall well-being in industrial settings. The goal is to create datasets that accurately represent the complexities found in real-world scenarios where physical fatigue notably influences human performance. These synthetically generated datasets will serve as the foundation for training specialized ML models designed explicitly for detecting the development of human physical fatigue. The trained ML model will undergo rigorous testing and validation using a substantial repository of authentic real-world data. The model's accuracy and reliability in detecting human physical fatigue will be assessed through this evaluation process. The ultimate objective is to achieve a level of accuracy that demonstrates the model's proficiency in identifying and predicting the onset of physical fatigue in human workers within industrial settings. This research endeavours to bridge the gap between Industry 4.0 innovations and human well-being by leveraging synthetic data generation techniques to enhance the accuracy and efficiency of ML models in detecting human physical fatigue.

Keywords: Synthetic Data Generation (SDG), Tabular Generative Adversarial Networks (GANs), Human Physical Fatigue Detection, Machine Learning (ML) Models

INTRODUCTION

Since the 1860s, researchers have studied fatigue, which is commonly characterised as a decline in mental and/or physical function brought on by factors such as mental strain, physical activity, inadequate sleep, interruption of the circadian rhythm or pattern, and sickness (Mohanavelu et al., 2017). Different scientific disciplines have offered ways to define and measure fatigue, but none of them apply universally to all areas. The main obstacles to providing a single definition of fatigue are, for example, its multidimensionality, the interaction of various variables (including confounding factors that must be accounted for), and, often, the subjective character of fatigue assessment (Marcus Yung, 2016).

Intelligent agent systems, sensing devices, and automation are becoming more sophisticated because of the Industrial Revolution (I5.0). Robotic systems and virtual assistance systems for work optimisation are being used more frequently in manufacturing and warehousing operations due to e.g. this rise in automation. This new I5.0 era is marked by its emphasis on highly skilled individuals who may profit from technological advancements (Iqbal et al., 2022). Due to the evolution of human-in-the-loop technology, collaborative robots (Cobots) have emerged and are being developed at a rapid pace. Despite their high level of automation, industries including aerospace, medical, pharmaceuticals, and manufacturing still involve highly fatiguing tasks. Workplace fatigue impacts a worker's productivity and is multifaceted. Even with Cobots being designed and deployed to alleviate human workload as well as augment human performance, the repetitive physical exertion demanded of daily tasks contributes to fatigue. Addressing fatigue in occupational health and safety is crucial due to its significant short- and long-term consequences. Regarding this paper, we aim to explore the complexity of fatigue, considering its multidimensionality and subjective assessment. Emphasizing the significance of synthetic data alongside real-time information from physiological sensors, our goal is to advance the understanding of fatigue for more effective interventions.

To address the issue of human fatigue within and across many industries, Machine Learning (ML) has been one of the most promising tools researchers employ to comprehend physical fatigue onset and development. This application involves leveraging ML algorithms to process and interpret data related to various physiological and behavioural parameters, aiming to identify signs of fatigue in individuals, amongst teams and so on.ML relies heavily on data for training, validating, and testing models. The quality and quantity of the data used play a crucial role in determining the performance and effectiveness of an ML model (Sedighi Maman et al., 2020). Obtaining human physical and behavioural data for machine learning algorithms can however pose several challenges. These challenges may include privacy concerns (Raghunathan, 2021), limited access to real data, data diversity and representativeness, bias (even partial) in real real-world data, limited availability of annotated data, and imbalance in social class distribution (Sedighi Maman et al., 2017).

In response, this paper explores the generation of synthetic tabular data and human physical fatigue datasets, addressing inherent challenges. It acknowledges the advantages of synthetic data, emphasizing cost-efficiency and ethical considerations while recognizing challenges such as complexity preservation and the need for domain-specific knowledge. The methodology introduces GAN for synthetic tabular data generation, emphasizing realism and privacy preservation. It

involves comprehensive data preprocessing for compatibility, including handling missing values and scaling features. Creating a human physical fatigue dataset through GANs entails training on real-world data. Considering the benefits of synthetic data, practitioners face a difficult landscape due to issues like complexity preservation and validation. Hence, a substantial amount of relevant research in the subsequent section is being conducted within a notable prevalence in the medical and financial sectors.

RELATED WORK

In recent years, the field of human physical fatigue detection has witnessed substantial advancements, with various methodologies and technologies being explored to monitor and analyse individuals' physiological and behavioural responses. These efforts largely aim to enhance occupational health and safety, optimize performance, and mitigate the potential risks associated with fatigue-related impairments. While a plethora of methods for fatigue detection exist, ranging from wearable sensors to machine learning algorithms, these approaches are not without their challenges (Lambay et al., 2022).

As aforementioned, the domain of human physical fatigue detection encounters challenges intrinsic to privacy concerns and these constraints pose impediments to robust model training, attributed to ethical considerations, regulatory frameworks, and proprietary constraints (Di Milia et al., 2011). Not only are the challenges pivotal, but they also extend to machine learning's reliance on extensive datasets. This dependence, however, introduces its own set of challenges, including the cost associated with sensors, time-consuming data acquisition processes, and issues such as missing values and misinformation (Lambay et al., 2021).

A promising alternative gaining traction is synthetic data generation. These methods create artificial datasets replicating real-world data statistics, addressing scarcity and privacy concerns (Li et al., 2021). Synthetic data generation has the potential to overcome limitations associated with traditional real data use, offering opportunities to advance human physical fatigue detection methodologies.

The significance of synthetic data in training robust and universal fatigue detection models cannot be overstated. Traditional datasets may lack the diversity and complexity required to capture the nuances of real-world scenarios in human-robot collaboration effectively such as considering the socio-demographics of a person. Regarding, generating synthetic datasets, the GAN model has been very popular among researchers as it has emerged as a promising solution to address this gap. GANs, are known for their ability to generate realistic, multimodal, multidimensional, and diverse datasets (Fonseca & Bacao, 2023). Recent studies, such as (Al-Qerem et al., 2023), have shown outcomes of an experimental investigation focused on enhancing the efficacy of SDG for multidimensional imbalanced datasets through the combinations of GAN models and Recursive Feature Elimination (RFE) technique. Similarly, aligned with the investigative approach undertaken by Saravana Kumar in 2017, this study systematically examined various GAN models to discern their respective efficacies. The empirical evidence substantiated the superior performance of PATECTGAN and CTGAN in this comparative analysis (Kiran & Saravana Kumar, 2017).

Nevertheless, the application of synthetic data generation has predominantly been noted in the context of generating patient or medical-related data (Abedi et al., 2022). This choice is informed by the inherent multidimensional, discrete, and

multimodal nature of medical data, which often exhibits a pronounced imbalance. For instance, one of the studies focuses on enhancing fluid overload prediction in intensive care units (ICUs) by integrating synthetic data with the existing medication dataset. Four ML algorithms were devised and trained on both the original and synthetically generated datasets, resulting in improved model performance (Rafiei et al., 2024). An additional study seeks to address challenges prevalent in applying machine learning to medical and cancer research, arising from issues such as data scarcity and privacy concerns. The authors systematically examine three classes of synthetic data generation, employing metrics to assess the quality of generated datasets derived from publicly accessible cancer registry data (Goncalves et al., 2020).

However, very limited research has focussed on generating human fatigue data which is also multidimensional and multimodal. One study by Lacasa et al. 2023, found the SDG techniques for creating chronic fatigue syndrome questionnaire datasets. However, in a more recent study, researchers attempted to apply SDG within an industrial scenario. In this study, they tried to achieve SDG by RGB image generation for human-object interaction (Leonardi et al., 2023). This shows the lack of available datasets and multimodal synthetic datasets for human physical fatigue in an industrial scenario that could train data-hungry models such as deep learning models. This paper introduces a robust methodology employing GAN for synthetic tabular data generation, addressing challenges in creating realistic and privacy-preserving datasets.

METHODOLOGY

Framework of SDG

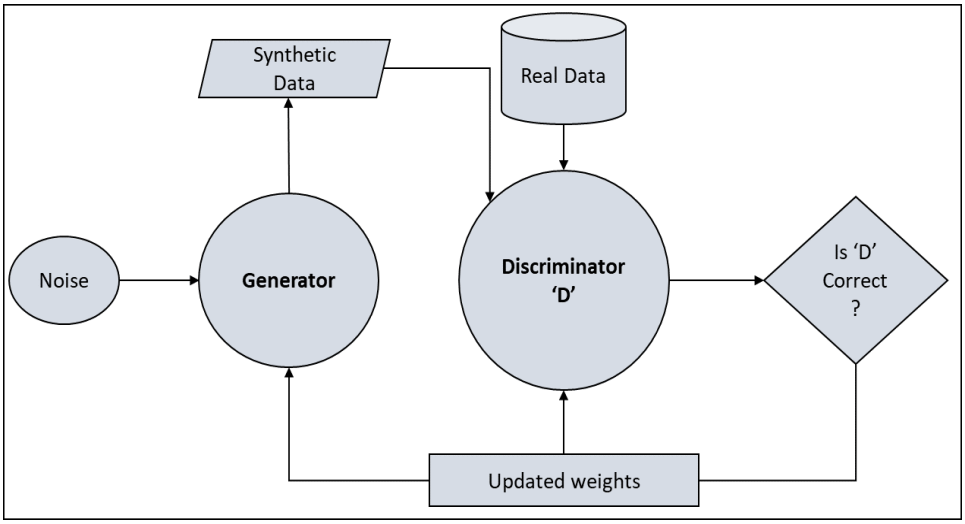

Figure 1: Framework of the GAN model used for synthetic data generation. Inspired by Li et al., 2021.

Figure 1 illustrates the general framework used to generate the human physical fatigue datasets. It consists of the framework that depicts different deep learning models to be used for SDG. In the deep learning models, the GAN architecture is pivotal to our methodology and comprises a generator and a discriminator network. The three main GAN learning models investigated for the current study are:1. Time-Series DGAN, 2. Tabular ACTGAN, and 3. Tabular LSTM GAN. These

adversarial networks were chosen for the training process due to their utilization of conditional GANs. These models specifically address challenges in controlling generated data and handling imbalanced tabular data by incorporating a conditional vector (Al-Qerem et al., 2023). The models refine the generator's ability to generate realistic tabular data to stabilize GAN training and enhance the quality of synthetic data, and so a training procedure is implemented. This procedure involves optimizing hyperparameters, incorporating regularization techniques, and continuously monitoring convergence. This is achieved by employing two models as shown in Figure 1, the Generator model and discriminator model.

The generator model employs a fixed-length random vector, drawn from a Gaussian distribution, to produce samples within a designated domain. This vector functions as the seed for the generative process and establishes a compressed representation known as a latent space, housing latent variables crucial for the domain. Conversely, the discriminator model assesses the authenticity of input examples, distinguishing between real and generated instances. Post-training, the generator, having acquired effective feature extraction capabilities, can be repurposed for its feature extraction layers for similar input data. To quantitatively analyse the performance of the generated synthetic data, we used Principal Component Analysis (PCA), correlation analysis and Field Distribution Stability. These metrics include measures which capture the dataset's diversity, and similarity indices to gauge the resemblance between the synthetic and original datasets.

Model Training for SDG

Data Gathering for Synthetic Data Generation (SDG): The initial phase involves acquiring pertinent data essential for Synthetic Data Generation (SDG). This data serves a dual purpose, being crucial for testing the generated data and training the model on both authentic and synthetic datasets. In this study, data has been sourced from open source, specifically focusing on the detection of human physical fatigue through inertial measurement units (IMU) and heart rate sensors (Lambay et al., 2022).

Data Preparation for SDG: The subsequent stage in SDG encompasses data preparation techniques. Tasks such as handling missing values and transforming data are imperative to address potential issues within the dataset.

Model Training (LSTM-GAN and ACTGAN): Following data preparation, the next step involves the training of selected models, namely LSTM-GAN and ACTGAN. Despite many similarities in their training approaches, distinctive differences characterize these models. The LSTM-GAN model features a structure consisting of two layers of LSTM generators, while the discriminator is a four-layer multilayer perceptron (MLP) utilizing the Adam optimizer and ReLU activation. Both generator and discriminator models undergo training across 5000 epochs, incorporating dropout layers to prevent overfitting.

GAN Training Process: For GAN models, beyond the typical training algorithm, a more monitored environment is necessary. This is achieved by looping over each epoch during training and selecting a random sample from the actual dataset. In the case of both models, adversarial training is employed for training various Generator and Discriminator models. Also, would investigate which optimization strategies can be employed to enhance the efficiency and convergence speed of the GAN training process.

Enhancements and Exploration of GAN Algorithms: The ACTGAN model, evolving from the well-established CTGAN model, introduces algorithmic enhancements that elevate effectiveness, precision, storage utilization, and conditioned generation capabilities. This comprehensive approach facilitated the exploration and evaluation of various GAN algorithms, with outcomes detailed in the subsequent section. An example question arising from this exploration is whether the integration of a synthetic dataset in an experimental study would impact the performance of machine learning algorithms.

RESULTS AND DISCUSSIONS

All GAN models underwent rigorous training and thorough analysis, following the methodology outlined, to assess the efficacy of synthetic data generation for human physical fatigue detection models. Given the well-established understanding that human fatigue is inherently subjective, the endeavour to capture diverse conditions and patterns of fatigue onset through biomechanical sensors is of paramount importance.

Figure 2:(A) Field Correlation Stability (B) Field Distribution Stability (C) Principal Component Analysis (PCA) of the Real(training) and Synthetic data of the Tabular LSTM-GAN Model respectively.

The nuanced nature of human fatigue necessitates a comprehensive and methodical approach to ensure that the generated synthetic data accurately reflects the multifaceted aspects associated with the manifestation of fatigue in varied contexts. Within the scope of our investigation, the performance hierarchy of the implemented models was discerned, with the time-series dGAN exhibiting the least efficacy, followed sequentially by ACTGAN. Foremost in performance was the LSTM-GAN, emerging as the most proficient among the evaluated models. This hierarchical evaluation offers valuable insights into the relative effectiveness of the considered models in the specific context of our study. Figure 2 illustrates the different evaluation analyses carried out on the Tabular LSTM-GAN model.

The images show that the field correlation stability, which is an initial step involving the computation of correlations between each pair of variables within both the training and synthetic datasets. The resulting average value serves as a quantitative indicator, wherein a diminished average signifies.

This methodology is encapsulated within the Field Correlation Stability quality score, where a lower average differential value corresponds to an elevated level of stability across the considered fields. In the context of LSTM, a noteworthy observation is made that its proficiency in faithfully replicating synthetic data to closely resemble the real dataset. The discernible aptitude of LSTM in accurately emulating the characteristics of real data underscores its efficacy as a generative model within the considered framework. The Field Distribution Stability, as depicted in image B of Figure 2, and further evidenced through PCA in image C, underscores a discernible resemblance in the distribution patterns when juxtaposed with the real dataset. This observation is indicative of a noteworthy coherence in the distributional characteristics between the synthetic and authentic datasets, reinforcing the fidelity of the synthetic data generation process in emulating the original dataset's inherent structures.

Figure 3: (A) Field Correlation Stability (B) Field Distribution Stability (C) Principal Component Analysis (PCA) of the Real(training) and Synthetic data of ACTGAN Model respectively.

In the examination of ACTGAN, a conspicuous observation emerges concerning its limited efficacy in faithfully replicating synthetic data to closely resemble to the real dataset. The discernible limitations in ACTGAN's ability to accurately emulate the characteristics of real data underscore potential shortcomings as a generative model within the considered framework. Notably, the Field Distribution Stability, as illustrated in image B of Figure 3 and further corroborated through PCA in image C, reveals a discernible disparity in the distribution patterns when contrasted with the real dataset. This observation suggests a notable incongruity in the distributional characteristics between the synthetic and authentic datasets, indicating potential challenges in maintaining

fidelity during the synthetic data generation process and capturing the inherent structures of the original dataset.

CONCLUSION

The novel study presented an investigation employing diverse Generative Adversarial Networks (GANs) to generate synthetic datasets for the training of human physical fatigue detection models. Among the triad of models utilized, the Tabular LSTM-GAN model demonstrated superior performance, notably exemplified by optimal Field Distribution Stability, an observation substantiated by Principal Component Analysis (PCA) and illustrated within Figures 2 and 3. Moreover, during the hyperparameter tuning process, it was observed that fine-tuning the model's hyperparameters, treated as optimization strategies, yielded optimal results for this model. The synthetic data generated by this model exhibited a pronounced concordance with the characteristics of the authentic dataset. Interestingly, the ACT-GAN model, an extension derived from the CT-GAN model, exhibited unforeseen efficacy, particularly in adeptly managing multimodal and multidimensional datasets intrinsic to fatigue datasets. However, it did not manifest a substantial impact on the performance of the machine-learning algorithm. Conversely, the Time Series dGAN model exhibited the least favourable outcomes, displaying suboptimal results and a notable disparity from the generation of datasets like the authentic example.

Subsequent research endeavours hold the potential to develop and investigate more advanced methodologies involving GAN for data augmentation and feature selection, thereby augmenting the existing knowledge base. The exploration of cutting-edge techniques and experimenting more with all the hyperparameters in these domains could offer heightened precision and efficacy in enhancing the performance of classifiers. Furthermore, extending investigations to encompass a diverse array of variables (biomechanical, gait, questionnaires, etc), classifiers and datasets would contribute valuable insights into the generalizability and adaptability of these SDG techniques across various contexts. By embracing a more expansive scope, future research could provide a comprehensive understanding of the broader applicability and replicability associated with advanced synthetic data generation and feature selection strategies.

REFERENCES

Abedi, M., Hempel, L., Sadeghi, S., & Kirsten, T. (2022). GAN-Based Approaches for Generating Structured Data in the Medical Domain. *Applied Sciences*, *12*(14), 7075. https://doi.org/10.3390/app12147075

Al-Qerem, A., Ali, A. M., Attar, H., Nashwan, S., Qi, L., Moghimi, M. K., & Solyman, A. (2023). Synthetic Generation of Multidimensional Data to Improve Classification Model Validity. *Journal of Data and Information Quality*, *15*(3). https://doi.org/10.1145/3603715

Di Milia, L., Smolensky, M. H., Costa, G., Howarth, H. D., Ohayon, M. M., & Philip, P. (2011). Demographic factors, fatigue, and driving accidents: An examination of the published literature. *Accident Analysis and Prevention*, *43*(2), 516–532. https://doi.org/10.1016/j.aap.2009.12.018

Fonseca, J., & Bacao, F. (2023). Tabular and latent space synthetic data generation: a literature review. *Journal of Big Data*, *10*(1), 115. https://doi.org/10.1186/s40537-023-00792-7

Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., & Sales, A. P. (2020). Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology*, *20*(1). https://doi.org/10.1186/s12874-020-00977-1

Iqbal, M., Lee, C. K. M., & Ren, J. Z. (2022). Industry 5.0: From Manufacturing Industry to Sustainable Society. *IEEE International Conference on Industrial Engineering and Engineering Management*, *2022-December*, 1416–1421. https://doi.org/10.1109/IEEM55944.2022.9989705

Kiran, & Saravana Kumar. (2017). *Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000. A methodology and an empirical analysis to determine the most suitable synthetic data generator*. *1*. https://doi.org/10.1109/ACCESS.2022.Doi

Lacasa, M., Prados, F., Alegre, J., & Casas-Roma, J. (2023). A synthetic data generation system for myalgic encephalomyelitis/chronic fatigue syndrome questionnaires. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-40364-6

Lambay, A., Liu, Y., Ji, Z., & Morgan, P. (2022). *Effects of Demographic Factors for Fatigue Detection in Manufacturing*. https://doi.org/10.1016/j.ifacol.2022.04.248

Lambay, A., Liu, Y., Morgan, P., & Ji, Z. (2021). A Data-Driven Fatigue Prediction using Recurrent Neural Networks. *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, 1–6. https://doi.org/10.1109/HORA52670.2021.9461377

Leonardi, R., Ragusa, F., Furnari, A., & Farinella, G. M. (2023). *Exploiting Multimodal Synthetic Data for Egocentric Human-Object Interaction Detection in an Industrial Scenario*. http://arxiv.org/abs/2306.12152

Li, D.-C., Chen, S.-C., Lin, Y.-S., Huang, K.-C., Huang, Y.-S. ;, Generative, K.-C. A., & Biancolillo, A. (2021). A Generative Adversarial Network Structure for Learning with Small Numerical Data Sets. *Applied Sciences 2021, Vol. 11, Page 10823*, *11*(22), 10823. https://doi.org/10.3390/APP112210823

Marcus Yung. (2016). *Fatigue at the Workplace: Measurement and Temporal Development*. University of Waterloo.

Mohanavelu, K., Lamshe, R., Poonguzhali, S., Adalarasu, K., & Jagannath, M. (2017). Assessment of Human Fatigue during Physical Performance using

Physiological Signals: A Review. *Biomedical and Pharmacology Journal*, *10*(4), 1887–1896. https://doi.org/10.13005/bpj/1308

Rafiei, A., Ghiasi Rad, M., Sikora, A., & Kamaleswaran, R. (2024). Improving mixed-integer temporal modeling by generating synthetic data using conditional generative adversarial networks: A case study of fluid overload prediction in the intensive care unit. *Computers in Biology and Medicine*, *168*. https://doi.org/10.1016/j.compbiomed.2023.107749

Raghunathan, T. E. (2021). Synthetic Data. *Annual Review of Statistics and Its Application*, *8*(1), 129–140. https://doi.org/10.1146/annurev-statistics-040720-031848

Sedighi Maman, Z., Alamdar Yazdi, M. A., Cavuoto, L. A., & Megahed, F. M. (2017). A data-driven approach to modeling physical fatigue in the workplace using wearable sensors. *Applied Ergonomics*, *65*, 515–529. https://doi.org/10.1016/J.APERGO.2017.02.001

Sedighi Maman, Z., Chen, Y. J., Baghdadi, A., Lombardo, S., Cavuoto, L. A., & Megahed, F. M. (2020). A data analytic framework for physical fatigue management using wearable sensors. *Expert Systems with Applications*, *155*. https://doi.org/10.1016/J.ESWA.2020.113405