Synthetic Patient Data Generation and Evaluation in Disease Prediction Using Small and Imbalanced Datasets

Antonio J. Rodriguez-Almeida, Himar Fabelo, Samuel Ortega, Alejandro Deniz, Francisco J. Balea-Fernandez, Eduardo Quevedo, *Member, IEEE,* Cristina Soguero-Ruiz, Ana M. Wägner and Gustavo M. Callico, *Senior Member, IEEE*

***Abstract*—The increasing prevalence of chronic non-communicable diseases makes it a priority to develop tools for enhancing their management. On this matter, Artificial Intelligence algorithms have proven to be successful in early diagnosis, prediction and analysis in the medical field. Nonetheless, two main issues arise when dealing with medical data: lack of high-fidelity datasets and maintenance of patient's privacy. To face these problems, different techniques of synthetic data generation have emerged as a possible solution. In this work, a framework based on synthetic data generation algorithms was developed. Eight medical datasets containing tabular data were used to test this framework. Three different statistical metrics were used to analyze the preservation of synthetic data integrity and six different synthetic data generation sizes were tested. Besides, the generated synthetic datasets were used to train four different supervised Machine Learning classifiers alone, and also combined with the real data. F1-score was used to evaluate classification performance. The main goal of this work is to assess the feasibility of the use of synthetic data generation in medical data in two ways: preservation of data integrity and maintenance of classification performance.**

***Index Terms* — Synthetic Data, Artificial Intelligence, Machine Learning, Generative Adversarial Networks, Classification, Data Augmentation, Imbalance**

This work was supported in part by the Spanish Government and European Union (FEDER funds) as part of support program in the context of TALENT-HExPERIA (HypErsPEctRal Imaging for Artificial intelligence applications) project, under contract PID2020-116417RB-C42, and by the project PID2019-107768RA-I00 (AAVis-BMR) and by the found action by the Community of Madrid in the framework of the Multiannual Agreement with Rey Juan Carlos University in line of action 1, "Encouragement of Young Phd students investigation"' Project Mapping-UCI (Ref F661). Moreover, this work was completed while Antonio Rodríguez was beneficiary of a pre-doctoral grant given by the "Agencia Canaria de Investigacion, Innovaci+on y Sociedad de la Información (ACIISI)" of the "Consejería de Economía, Conocimiento y Empleo", which is part-financed by the European Social Fund (FSE) (POC 2014-2020, Eje 3 Tema Prioritario 74 (85%)) and, Himar Fabelo was beneficiary of the FJC2020-043474-I funded by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU/PRTR". *(Corresponding author: Himar Fabelo).*

A. J. R.-A., H. F., S. O., F. J. B.-F., E. Q. and G. M. C. are with the Research Institute for Applied Microelectronics, University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain (e-mail: aralmeida@iuma.ulpgc.es; hfabelo@iuma.ulpgc.es; sortega@iuma.ulpgc.es; fbalea@cop.es; equevedo@iuma.ulpgc.es; gustavo@iuma.ulpgc.es).

S. O. is also with Norwegian Institute of Food, Fisheries and Aquaculture Research, Tromsø, Norway.

H.F. is also with Fundación Instituto de Investigación Sanitaria de Canarias, Las Palmas de Gran Canaria, Spain.

F. J. B.-F.is also with Dept. of Psychology, Sociology and Social Work, University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain.

A. D. and A. M. W. are with the Endocrinology and Nutrition Department, Complejo Hospitalario Universitario Insular Materno-Infantil and with the Institute of Biomedical and Health Research. University of Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain (e-mail: aledenizgarcia@gmail.com; ana.wagner@ulpgc.es).

C. S.-R. is with the Department of Signal Theory and Communications, Telematics and Computing Systems, Rey Juan Carlos University, Fuenlabrada, Madrid, Spain (e-mail: cristina.soguero@urjc.es).

The codes developed in the current study are available from the corresponding author on reasonable request. Code is also publicly available on the following GitHub repository: (github.com/antorguez95/synthetic_data_generation_framework).

I. Introduction

ARTIFICIAL Intelligence (AI) is arising as a potential truly helpful tool in clinical practice. Risk factor analysis, disease or disorder prediction, risk estimation or image segmentation are just a few examples of how AI could help physicians in their work [1],[2]. Besides, there are some diseases that specially need tools to enhance their management due to their continuously increase in prevalence all over the world [3].

Even though AI techniques are potentially beneficial in healthcare, they need to be fed with large amounts of data. When using small databases, these algorithms are not able to generalize to the overall population, and they overfit the data they are trained on [4]. In addition, there are two main issues that arise when dealing with medical data: (i) lack of available high-fidelity datasets and (ii) maintenance of patient's data privacy [5]. The former is due to the high cost of acquiring medical data and ethical bureaucracy associated to it, apart from the fact that some institutions might be suspicious to share their data. The latter refers to the potential, inappropriate re-identification of patients, even when using anonymized data, as patient's privacy must be kept [6].

To face these problems, different synthetic data generation techniques have emerged. Synthetic data generation can be defined as the technique of creating "fake" samples from a

[63] "Statistical functions (scipy.stats) — SciPy v1.8.0 Manual." https://docs.scipy.org/doc/scipy/reference/stats.html (accessed Feb. 15, 2022).

[64] F. Wang, R. Kaushal, and D. Khullar, "Should health care demand interpretable artificial intelligence or accept 'black Box' Medicine?," *Annals of Internal Medicine*, vol. 172, no. 1, pp. 59–61, Jan. 2020, doi: 10.7326/M19-2548.

[65] B. Vega-Márquez, C. Rubio-Escudero, and I. Nepomuceno-Chamorro, "Generation of Synthetic Data with Conditional Generative Adversarial Networks," *Logic Journal of the IGPL*, Nov. 2020, doi: 10.1093/jigpal/jzaa059.

[66] C. A. Libbi, J. Trienes, D. Trieschnigg, and C. Seifert, "Generating Synthetic Training Data for Supervised De-Identification of Electronic Health Records," *Future Internet 2021, Vol. 13, Page 136*, vol. 13, no. 5, p. 136, May 2021, doi: 10.3390/FI13050136.

[67] "Hide-and-Seek Privacy Challenge: Synthetic Data Generation vs. Patient Re-identification." https://proceedings.mlr.press/v133/jordon21a (accessed Jun. 15, 2022).

[68] "Privacy Preserving Synthetic Health Data - Inria." https://hal.inria.fr/hal-02160496/ (accessed Jun. 15, 2022).