Combining Synthetic Patient Data Generation with Machine Learning Methods for Diabetes Prediction

Antonio J. Rodriguez-Almeida[a*], María Castro-Fernandez[a], Alejandro Déniz[b], Himar Fabelo[a], Samuel Ortega[a], Eduardo Quevedo[a], Cristina Soguero-Ruiz[c], Ana M Wägner[b], Conceição Granja[d] and Gustavo M. Callico[a]

[a]Institute for Applied Microelectronics, University of Las Palmas de Gran Canaria (ULPGC), Spain.
[b]Endocrinology and Nutrition department, Complejo Hospitalario Universitario Insular Materno-Infantil de Gran Canaria, Spain. Research Institute of Biomedical and Health Sciences (IUIBS), ULPGC, Spain
[c]Dept of Signal Theory and Communications, Telematics and Computing Systems. Rey Juan Carlos University Fuenlabrada, Madrid, Spain
[d]Faculty of Nursing and Health Sciences, Nord University, Bodø, Norway

*E-mail: aralmeida@iuma.ulpgc.es

ABSTRACT

Diabetes Mellitus (DM) is a chronic disease caused by different disorders in the insulin production or use. Its prevalence has not stopped increasing during the last years, becoming a major public health concern. Thus, tools for its prediction and early diagnosis are needed. In this context, Machine Learning (ML) could be a suitable choice due to its capability of extracting useful information from medical records. However, the lack of available and reliable datasets makes this a complex task. Synthetic data generation is emerging as a solution for this issue, as it takes a real dataset as the basis to generate similar instances. In this work, a framework based on ML and synthetic data generation methods is presented to evaluate whether classification performance between presence or absence of DM could be improved. The obtained results show that ADASYN and Borderline SMOTE algorithms fairly keep the underlying structure of the original data. They also prove that the ML models trained with mixed synthetic and original data perform as well as those trained with original data.

Keywords: Machine Learning, Synthetic Data Generation, Diabetes Mellitus, Classification.

1. INTRODUCTION

Diabetes Mellitus (DM) is a chronic disease caused by either insufficient insulin production in the pancreas (insulin deficiency) and/or when the body is unable to properly use the insulin it produces (insulin resistance), leading to increased blood sugar. In diabetes, high blood sugar (hyperglycaemia) can lead to damage in nerves and blood vessels. According to the World Health Organization (WHO), in 2014, there were 422 million people with diabetes and its prevalence was, and still is, on the rise [1]. In 2019, over 1.5 million deaths worldwide were caused by diabetes, being a major cause of blindness, kidney failure, heart attack, stroke, and limb amputation [1]. Given its spread all across the world and the burden it poses on individuals and society alike, tools for its prediction and early diagnosis are needed.

Artificial Intelligence (AI) is an exponentially growing field. Within AI, Machine Learning (ML) can be defined as computer programming using data techniques, rather than classic algorithms. This technique is based on traditional statistics theory to build different types of mathematical models. ML algorithms are capable of extracting information from rich and complex datasets. ML combined with smartphones, wearables and medical devices could be a suitable tool help physicians, patients or other users in the prevention and management of chronic diseases. In the case of DM, ML could improve risk-prediction and diagnostic models, and treatment personalization, including self-management of the disease [2]. Many ML techniques have been applied to the study of DM and no technique is inherently better than the rest, although some are more frequently used to differentiate among DM types (i.e., Type 1, Type 2 or Gestational DM). On the other hand, applying a well-sized and balanced dataset is essential, because training an algorithm with an imbalanced dataset could lead to biased models, since ML algorithms may tend to ignore the minority classes [3]. In the same way, missing values lower the algorithm's ability to learn and even if it is not possible to avoid lack of information, arrangements could be made to facilitate the learning process [4].

Suitable open medical datasets are scarce and often not consistent, replicable, reliable, or big enough for study purposes. The main goal of this work is to evaluate if the application of data balancing and data augmentation methods, combined with ML algorithms, could lead to a more accurate DM prediction. Therefore, this work is presented as a preliminary approximation to a possible solution to this issue. A framework based on synthetic data generation and evaluation has been developed to classify between presence or absence of DM. The synthetic data generation was performed using different, validated methods for imputation, data balancing and data augmentation. This work intends to investigate if the methods do not significantly change the underlying structure of the original dataset, and to investigate whether the classification performance is not worsened by the use of synthetic data.

2. MATERIALS AND METHODS

2.1 PIMA Indians Diabetes database

The Pima Indians of Arizona, U.S.A, had the highest prevalence of DM worldwide [5]. The National Institute of Diabetes and Digestive and Kidney Diseases recorded data for females of Pima Indian heritage who were at least 21 years old [5]. In particular, the dataset was composed of a total of 768 individuals, 268 with DM and 500 without. The goal of this study is to predict whether or not a patient has diabetes based on eight independent variables: the number of pregnancies the patient has had; the body mass index ($weight[kg]/(height[m])^2$); the 2-hours plasma glucose

concentration in an oral glucose tolerance test; the diastolic blood pressure (mm Hg); the triceps skin fold thickness (mm); the 2-Hour serum insulin (mu U/ml); the diabetes pedigree function and the age (years). This dataset can be considered as a "worst case" scenario since there are features, such as insulin, whose values are missing in half of the whole dataset instances. Besides, it has been assumed that zero insulin values mean missing data, when, in reality, insulin dose can be zero. However, this information is not provided in the dataset.

2.2 Synthetic data generation techniques

Two different synthetic data generation techniques were used. Firstly, algorithms designed to balance imbalanced datasets were applied. Once the proportion of controls and cases was well-adjusted, algorithms to augment data from the whole dataset were used. We expected that data balance would avoid introducing bias in the model due to the existence of a majority class, while data augmentation would improve the generalization of the model.

2.2.1 Data Balancing

To balance this dataset, two widely used algorithms were applied [6]: Synthetic Minority Over-sampling Technique (SMOTE) and ADAptive SYNthetic Sampling (ADASYN). The basis of the SMOTE algorithm is to oversample the minority class introducing random samples along the line segments joining any (or all) *k* minority sample neighbours. Apart from the original SMOTE implementation, three additional variants of the original algorithm were tested. The *K-Means SMOTE* applies a K-Means clustering before oversampling with SMOTE [6]. The *SVM (Support Vector Machine) SMOTE* detects samples to use as a reference through a SVM classifier prior to oversampling [6]. The Borderline SMOTE (BS) detects the borderline samples of each class and only the minority samples near the borderline are oversampled [6]. The ADASYN algorithm can be considered as an improvement of the SMOTE algorithm. Whereas the SMOTE algorithm generates arbitrary minority examples, ADASYN uses weighted distributions for different minority class samples. This algorithm focuses on generating more samples from the actual samples that are the most difficult to learn [6].

2.2.2 Data Augmentation

As aforementioned, after data balance is performed, data augmentation is assessed. To that end, two different algorithms have been implemented: Gaussian Copulas (GCs) [7] and Conditional Tabular Generative Adversarial Networks (CTGANs) [8]. GCs are constructed from a multivariate normal distribution, being capable of reproducing a large variety of multivariate distributions. Generative Adversarial Networks (GANs) have been extensively used lately, and many variants have been developed with different purposes. In this case, since this dataset can be considered tabular data, CTGANs have been selected. The GANs are Deep Learning (DL) models based on a discriminative model that learns to determine if a sample belongs to the data distribution or to the model distribution. The generative model creates data that are evaluated by the discriminative model, so both improve their methods until generated data and real data are indistinguishable. CTGANs are GANs specifically designed to model tabular data, prepared to overcome the non-Gaussian and multimodal distributions of imbalanced datasets.

2.3 Machine learning techniques

For the classification task, four different supervised classifiers were used for performance comparison, and to test if mixing the synthetic data with the real data for training, worsened, improved, or kept such performances. The chosen classifiers were SVM [9], Random Forest (RF) [9], K-Nearest Neighbours (KNN) [9] and XGBoost (XGB) which is a gradient boosting algorithm [10]. A grid search was performed for all cases to find the optimal hyperparameters among those considered most relevant for each model.

2.4 Evaluation metrics

One of the main objectives of this work is to determine if the different synthetic data generation algorithms faithfully replicate the original dataset, or if, in contrast, the underlying structure of the data is totally or partially lost. Based on the literature regarding generation and evaluation of synthetic medical data, three different metrics have been used: 1) Pairwise Correlation Difference (PCD) describes how much correlation the synthetic data has been able to capture from the original data and is computed at the dataset level. The smaller the PCD, the closer the synthetic data are to real data in terms of linear correlations across the variables [11]; 2) Maximum Mean Discrepancy (MMD) indicates how well the model captures the distribution of the real data at dataset level. Lower MMD indicates higher similarity distribution. A value equal to zero means that the distribution was perfectly captured. It has been proven to be effective evaluating GANs [12]; 3) Kullback-Leibler Divergence (KLD) measures how different a probability distribution is from the original one using logarithmic functions. This parameter is computed at feature level, not at dataset level. Following, the KLD of every feature was added so a final single number was obtained. A value equal to zero means that distributions contain the same information. The higher the value, the worse the similarity between datasets. It is worth noticing that KLD does not measure dependencies among variables [11]. Since KLD matrices must have the same dimensions to be comparable, this metric is not used to evaluate balancing algorithms. Classification performance was evaluated in terms of accuracy (ACC), Area Under the Curve (AUC), and F1-score. Notice that accuracy could be high when training with the original dataset since this dataset has not been balanced. This result does not mean that the model learned from data is better than others, but means that it is biased by the majority class. For this reason, F1-score should be the main parameter to consider when evaluating these models.

2.5 Proposed processing framework

Figure 1 shows the data processing framework proposed in this work. This framework has been entirely developed in Python programming language, using `sklearn`, `imblearn` and `sdv` libraries for the ML and synthetic data generation processing. The first step is a manual cleaning of the raw dataset based on the prior knowledge of the physiological variables. With this, mistakes on the data acquisition and illogical values are removed. Then, the dataset is partitioned into training (80%) and test (20%) sets. Afterwards, missing data of both subsets are imputed using a KNN imputation algorithm. The test set was imputed using train data instances [24]. From here, the processing framework is only applied on the training set, while the test set is used on the final step. Once imputation is performed, data balancing algorithms are applied, so that the problems associated to model training with an imbalanced dataset are removed. At this point, the first evaluation of the synthetic data is assessed. Only the algorithms that show the best results in terms of statistical data similarity are passed through the data augmentation algorithms. In this step, data are augmented following two paths: a) with the entire balanced dataset, and b) cases and controls separately. Once augmented, synthetic data are again evaluated to check which algorithm replicates the underlying structure of the previous dataset in a more reliable way. Both the training and the test sets were separately normalized by centring and scaling the samples with the mean and standard deviation of each feature. Notice that the reference dataset in this evaluation will vary depending on the balance algorithm previously used. The final step consists of training the ML models (including their optimization following a 10-fold cross validation approach) with the processed dataset and validating them with the test set, evaluating and analysing the obtained results.

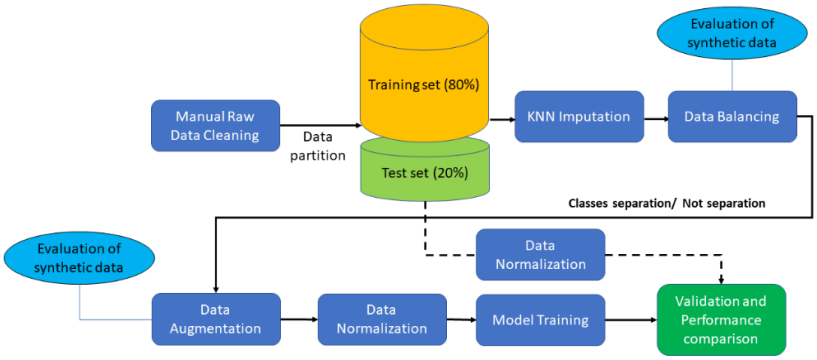


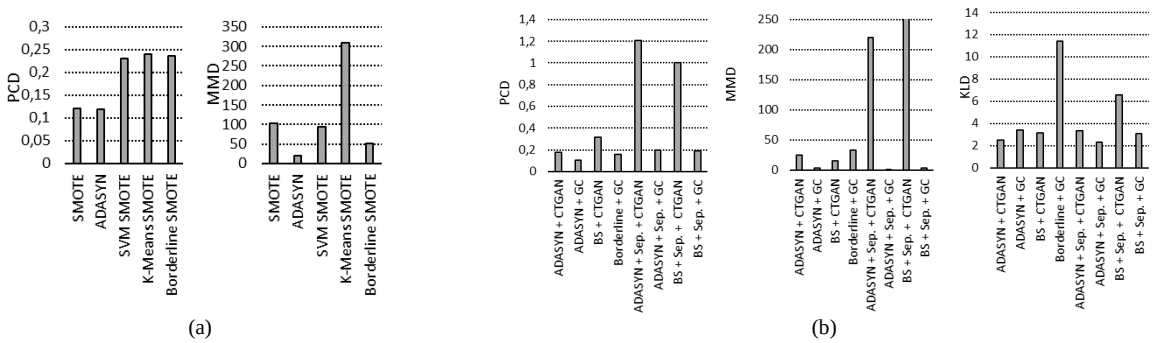
Figure 1. Proposed processing framework.

3. EXPERIMENTAL RESULTS

3.1 Synthetic data generation results

Depending on the data balancing algorithm used, different dataset dimension datasets were generated. The output of the ADASYN algorithm were 809 instances, while BS algorithm output were 800 instances. In the data augmentation step, CTGAN and GC always duplicated the instances (ending with 1618 and 1600, depending on the previous balancing algorithm). Hence, classifiers were trained with ~1600 instances, whereas without synthetic data, classifier were trained with 611 instances.

Figure 2.a illustrates the metrics that describe the quality of the data generated after balancing algorithms were applied to the imputed dataset. Notice that, when data balance algorithms are applied, it is usual to partially lose the underlying structure of the data. The Pearson correlation hardly changes, which means that the linear correlations have been fairly maintained. The balancing algorithms that best fit the original data were ADASYN and the BS, even though BS shows a slightly high value of PCD. Their output were the ones passed through the GC and the CTGAN algorithms and taken as the reference datasets to study the different synthetic data generation metrics. When the dataset is partitioned in control and cases after data balance, this is indicated with the word "Sep".



(a) (b)

Figure 2. Results of the different metrics obtained for data balance (a) and data augmentation (b) methods.

Figure 2.b shows the results after data augmentation was performed. According to the PCD, the GC algorithm always preserves the linear correlations much better than the CTGAN. Regarding the rest of parameters, one algorithm does

not clearly overcome the other. Based on these three parameters, the methods that, in general, preserve this diabetes dataset structure best are **ADASYN+GC**, **ADASYN+Sep+GC** and **BS+Sep+GC**.

3.2 Classification results

Table 1 shows the classification performance for the eight different combined methods and the four classifiers. The models trained with the original dataset without the presence of synthetic data are referred to as "reference". In general, the performance of the classifiers is similar regarding to the reference results. F1-Score reveals that training the classifiers using balanced and augmented synthetic data provides higher performance than using the original unbalanced dataset. In ACC and AUC metrics, the results are fairly similar to the reference, except for the SVM using BS+GC that improves such reference results. It can be noticed that RF and XGB classifiers achieve the highest performance.

Table 1. Classification results obtained using the four different classifiers.

	ACC				**AUC**				**F1-Score**			
	SVM	**RF**	**XGB**	**KNN**	**SVM**	**RF**	**XGB**	**KNN**	**SVM**	**RF**	**XGB**	**KNN**
Reference	0.77	**0.82**	**0.83**	**0.80**	0.87	**0.90**	**0.90**	**0.88**	0.64	0.72	0.72	**0.69**
ADASYN+CTGAN	0.75	0.72	0.75	0.67	0.86	0.85	0.88	0.82	0.72	0.66	0.71	0.64
ADASYN+GC	0.76	0.77	0.72	0.69	0.86	0.86	0.80	0.82	0.71	0.72	0.65	0.65
BS+CTGAN	0.75	0.75	0.76	0.72	0.83	0.85	0.84	0.85	0.70	0.67	0.69	0.67
BS+GC	**0.79**	0.75	0.75	0.74	**0.88**	0.88	0.87	0.87	**0.75**	0.70	0.68	**0.69**
ADASYN+Sep+CTGAN	0.68	0.71	0.69	0.66	0.75	0.77	0.70	0.68	0.57	0.62	0.55	0.56
ADASYN+Sep+GC	0.78	0.78	0.74	0.67	0.85	0.87	0.86	0.80	0.74	**0.73**	0.71	0.63
BS+Sep+CTGAN	0.73	0.76	0.66	0.65	0.78	0.81	0.72	0.72	0.64	0.66	0.52	0.60
BS+Sep+GC	0.77	0.75	0.78	0.69	0.85	0.87	0.88	0.80	0.73	0.69	**0.73**	0.65

4. CONCLUSIONS

In this work, it has been demonstrated that a framework that combines synthetic data generation with ML classification applied to a diabetes dataset does not worsen the classification performances, improving them in some cases in terms of F1-score metric. Furthermore, the proposed framework could offer robustness against overfitting and better generalization, since the models are trained using higher number of samples that reliably represents the original dataset. From the obtained results, it is clear that there is not any synthetic data generation algorithm that perfectly suits all classifiers. In general, better metrics in the synthetic data similarity implies better classification performance. The fact that this framework works well with this limited dataset is promising. In this sense, further work will continue with other datasets to validate the proposed approach to improve classification or apply it on other tasks such as regression or clustering.

5. ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 10101738, and from 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. Additionally, this work was completed while Antonio Rodríguez was beneficiary of a pre-doctoral fellowship by the "Agencia Canaria de Investigación, Innovación y Sociedad de la Información (ACIISI)" of the "Consejería de Economía, Conocimiento y Empleo" of the "Gobierno de Canarias", which is part-financed by the European Social Fund (FSE) (POC 2014-2020, Eje 3 Tema Prioritario 74 (85%)).

REFERENCES

1. WHO. Diabetes [Internet]. [cited 2021 Sep 28]. Available from: https://www.who.int/news-room/fact-sheets/detail/diabetes
2. Chaki J, Thillai Ganesh S, Cidham SK, Ananda Theertan S. Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. J. King Saud Univ. - Comput. Inf. Sci. King Saud bin Abdulaziz University; 2020.
3. Sharma S, Bellinger C, Krawczyk B, Zaiane O, Japkowicz N. Synthetic Oversampling with the Majority Class: A New Perspective on Handling Extreme Imbalance. Proc - IEEE Int Conf Data Mining, ICDM. Institute of Electrical and Electronics Engineers Inc.; 2018;2018-November:447–56.
4. García S, Luengo J, Herrera F. Dealing with Missing Values. Intell Syst Ref Libr. Springer, Cham; 2015;72:59–105.
5. Bennett P, Burch T, Miller M. Diabetes Mellitus in American (PIMA) Indians. Lancet. Elsevier; 1971;298:125–8.
6. Babar VS, Ade R. A Review on Imbalanced Learning Methods. IJCA Proc Natl Conf Adv Comput. 2015. p. 23–7.
7. Xue-Kun Song P. Multivariate Dispersion Models Generated From Gaussian Copula. Scand J Stat. 2000;27:305–20.
8. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling Tabular Data using Conditional GAN.
9. Balea-Fernandez FJ, Martinez-Vega B, Ortega S, Fabelo H, Leon R, Callico GM, et al. Analysis of Risk Factors in Dementia through Machine Learning. J Alzheimer's Dis. 2021;79.
10. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System.
11. Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. BMC Med Res Methodol. BioMed Central; 2020;20:108.
12. Sutherland DJ, Tung H-Y, Strathmann H, De S, Ramdas A, Smola A, et al. Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy. 5th Int Conf Learn Represent ICLR 2017 - Conf Track Proc. International Conference on Learning Representations, ICLR; 2016;