A novel missing data imputation approach based on clinical conditional Generative Adversarial Networks applied to EHR datasets

Michele Bernardini, Anastasiia Doinychko, Luca Romeo, Emanuele Frontoni, Massih-Reza Amini

A novel missing data imputation approach based on clinical conditional Generative Adversarial Networks applied to EHR datasets

Michele Bernardini[a,*], Anastasiia Doinychko[d], Luca Romeo[b], Emanuele Frontoni[c], Massih-Reza Amini[d]

[a]*Department of Information Engineering (DII), Università Politecnica delle Marche, Ancona, Italy*
[b]*Department of Economics and Law, University of Macerata, Macerata, Italy*
[c]*Department of Political Sciences, Communication and International Relations, University of Macerata, Macerata, Italy*
[d]*Grenoble Informatics Laboratory, Université Grenoble Alpes, Saint-Martin-d'Hères, France*

Abstract

The missing data mechanism is a relevant problem in Machine Learning (ML) and biomedical informatics communities. Real-world Electronic Health Record (EHR) datasets comprise several missing values, thus revealing a high level of spatiotemporal sparsity in the predictors' matrix. Several approaches in the state-of-the-art tried to deal with this problem by proposing different data imputation strategies that (i) are often unrelated to the ML model, (ii) are not conceived for EHR data where laboratory exams are not prescribed uniformly over time and percentage of missing values is high (iii) exploit only univariate and linear information on the observed features. Our paper proposes a data imputation strategy based on a clinical conditional Generative Adversarial Network (ccGAN) capable of imputing missing values by exploiting non-linear and multivariate information across patients. Unlike other GAN data imputation-based approaches, our method deals explicitly with the high level of missingness of routine EHR data by conditioning the imputing strategy to the observable values and those fully-annotated. We demonstrated the statistical significance of the ccGAN to other state-of-the-art approaches in terms of imputation and predictive performance on a real

*Corresponding author: phone +390712204458, fax +390712204224
Email addresses: m.bernardini@pm.univpm.it (Michele Bernardini), anastasiia.doinychko@univ-grenoble-alpes.fr (Anastasiia Doinychko), luca.romeo@unimc.it (Luca Romeo), emanuele.frontoni@unimc.it (Emanuele Frontoni), massih-reza.amini@univ-grenoble-alpes.fr (Massih-Reza Amini)

multi-diabetic centers dataset. We measured its robustness across different missingness rates on an additional benchmark EHR dataset.

Keywords:
Data imputation, Generative Adversarial Network, Electronic Health Record, Machine Learning, Predictive Medicine.

---

1. Introduction

Given the increasing and unavoidable digital transformation process of national healthcare system management, the considerable size of structured Electronic Health Record (EHR) data is becoming available.

In predictive and precision medicine, Machine Learning (ML) techniques can manage EHR data and provide disease predictions [**?**]. On the other hand, ML's potential may be limited by the low quality of the EHR data, i.e. high sparsity, imbalanced setting, noisy and redundant features, and irregular time sampling characteristics. This challenging scenario is emphasized in routine EHR data (i.e., general practitioners, diabetic centers, and clinics), where not all laboratory exams are prescribed uniformly over time. This scenario contrasts the Intensive Care Unit (ICU) setting, where the exams are performed regularly over time, thus leading to a much more complete and uniform data collection. For these reasons, an adequate and effective missing data imputation stage is crucial in the data preprocessing pipeline. Specifically, a suitable data imputation strategy may positively influence the effectiveness of the ML algorithm for prognosis and disease prediction. The missing data mechanism can be categorized into i) completely at random, ii) at random, or iii) not a random [**?**]. In our case, we provide results under the missingness completely at random (MCAR) assumption. Moreover, experimental results are also provided under the real-clinical scenario where laboratory exams are not prescribed uniformly over time.

This study seeks to offer a data imputation technique based on a clinical conditional Generative Adversarial Network (ccGAN) capable of imputing missing values of observed characteristics conditioned by fully-available characteristics values to be then employed for predicting the probable diabetes complication.

We investigated our proposed strategy via the lens of a specific clinical use case (i.e., diabetic retinopathy (DR) prediction) of diabetes complications by using a real EHR multi-diabetic centers (MDC) dataset. We evaluated the performance comparison of commonly used imputation algorithms for the MDC dataset to accommodate high missingness rates (up to 80% for the whole MDC dataset and up to 40% per patient). DR caused by chronically high or variable blood sugar is the most typical and insidious diabetes microvascular complication. With the worldwide increasing incidence of diabetic patients with DR and consequential visual impairments, early diagnosis and timely appropriate treatment are progressively becoming effective measures to prevent DR and alleviate the economic burden over the national healthcare systems [**?**]. Physicians typically diagnose the DR by directly evaluating fundus images, but this gold standard process, usually carried out when the DR has already been delineated, remains expensive, time-consuming, and sometimes unnecessary [**?**]. Thus, the early prediction of developing DR by employing only routine EHR data and Machine Learning (ML) techniques may result in a convenient and effective strategy for follow-up diabetic patients within a screening scenario. Then, in the case of a possible positive DR prediction, the next complementary - but not alternative - step would be to perform a gold standard diagnosis to confirm/deny the predictive model output. Following this rationale (i.e., favorable clinical implications in managing diabetic patients), the proposed strategy can be easily extended and generalized to other common diabetes complications (e.g., nephropathy, cardiopathy, etc.) or other predictive clinical tasks. Therefore, through an additional experiment on a widely employed EHR ICU benchmark dataset (MIMIC-III dataset [**?**]), we also measured the imputation and predictive performance of the proposed ccGAN approach across different missingness rates.

The main contributions to biomedical informatics are threefold and can be summarized as follows:

- we proposed an ML approach to impute missing values from EHR data and provide the prediction of DR. The data imputation strategy is based on a novel ccGAN architecture that exploits the fully-available clinical features among different patients to infer other missing clinical features. The prediction phase is

realized by implementing and comparing different ML classifiers;

- we showed how the proposed ccGAN approach overcomes significantly other state-of-the-art methodologies in terms of data imputation and predictive performance to solve DR tasks using a real MDC dataset.

- we demonstrated the robustness of our proposed approach across different missingness rates in terms of both data imputation and predictive performance on an additional benchmark EHR dataset.

The rest of the manuscript is organized as follows: Section 2 provides an overview of the state-of-the-art data imputation techniques in EHR data; Section 3 describes the employed real EHR dataset (i.e., MDC dataset) and the preprocessing procedure; Section 4 describes the proposed method, the experimental procedure, and comparisons; Section 5 shows the imputation and predictive performance results; Section 6 discusses the experimental findings and Section 7 concludes the paper.

2. Related work

We review the work on data imputation techniques tailored to imputing missing EHR data. We decided not to consider strategies that tried to model the temporal dependencies among missing values. This assumption excludes all methods of applying recurrent neural network-based imputation methods for modeling sequential patterns and dependencies in time series. Although these approaches assume potential relevance for ICU data, they cannot be applicable for treating routine EHR data as longitudinal time series [**? ?**]. The rationale motivation is that the amount of observations per patient in our real clinical MDC dataset is limited and sparsely distributed over time (i.e., on average 19 observations per patient). In this particular setting, a non-temporal correlation among univariate characteristics is expected, as confirmed in Section 4.1.

Case deletion methods (i.e., instances with missing elements are removed) are among the most straightforward approaches that may potentially lose some valuable information in EHR data [**?**]. Instead of just dropping patients with missing elements, a more suitable strategy would be to replace the missing values. EHR data may be

intentionally missing (e.g., the patient does not need such laboratory tests) or unintentional (e.g., lack of routine check-ups or follow-ups). However, if not considered, this missingness might result in a loss of power, biased estimates, and underperformed models. [**?**]. Notably, imputing missing real EHR data (e.g., irregular time series) is a persistent challenge. An optimal standardized imputation strategy given a defined missingness pattern has yet to be proposed in the literature. Various traditional imputation methods have been successfully applied, including mean, median, and extra value substitution. Also, traditional statistical methods (i.e., expectation maximization, full information maximum likelihood, multiple imputations) [**?**] and interpolation methods (i.e., linear, polynomial, backward/forward, padding) have been largely adopted to impute missing values in EHR data. Multivariate Imputation by Chained Equations (MICE) [**?**] adopts a chained equation over various iterations to estimate the missing values after an arbitrary initialization. Collaborative filtering approaches based on expectation-maximization and matrix-factorization [**?**] represent the patient in a lower-dimensional latent space to impute missing information. The major limitation of these approaches is that they deal with a low percentage of missing values; thus, the imputation accuracy decreases as quickly as the percentage increases. This drawback originates because these strategies only sometimes succeed in capturing non-linear relationships between observed and unobserved features.

In ML-based imputation approaches, the imputation and prediction tasks are usually performed asynchronously (i.e., the prediction task is separate from the data imputation mechanism) or simultaneously (i.e., a unique model encloses both data imputation and prediction tasks) [**?**]. Some ML approaches require a part of complete data to fill the missing values [**?**], while others also take into account partial missing training data [**?**], [**?**]. ML-based imputation models like K-Nearest Neighbors (KNN) [**?**] and MissForest (MissF) [**?**] are also among the most common methods. However, KNN requires tuning of the parameter k and is vulnerable and sensitive to outliers. Additionally, KNN may be computationally expensive. MissF, based on the Random Forest (RF) algorithm, is robust to noisy data and multicollinearity since random forests have built-in feature selection. However, MissF does not consider any multivariate information among features to capture the missing mechanism. Addition-

ally, many trees in the MissF may slow down and make the data imputation process ineffective for real-time EHR data. Unlike other ML-based imputation algorithms that usually exploit univariate information (KNN, MissF) or linear information (MICE, matrix factorization) across features, our approach allows computing the missing value of the candidate feature based not only on the available value of the selected feature but also based on the value of other features. Thus, we consider GAN-based imputation methods that (i) are successfully applied with complex data distribution (e.g., collection of images, texts, time series, or those similar to the actual observed EHR data); (ii) employ neural networks that allow exploiting any non-linear relation.

GANs are state-of-the-art solutions to distribution modeling tasks defined by a collection of data of any complexity. In the direction of problems with missing values, first, advanced GAN-based models are proven to recover a distribution of interest from lossy observations only [**?**], [**?**]. Secondly, recent algorithms propose data imputation strategies in different real-world data domains of study, such as incomplete image imputation - MisGAN [**?**]; medical records data imputation - GAIN [**?**]; and missing view generation - CollaGAN [**?**], VIGAN [**?**]. Notably, the last two models tackle the missing view problem in multi-domain datasets, which is challenging. However, the main interest of particular work is learning with partially observed data - the same objective as for MisGAN and GAIN algorithms. While both provide affordable results for missing data imputation on data with low and high rates of incompleteness, MisGAN architecture, in its turn, employs six neural networks in total to train, which makes it computationally expensive and time-consuming to work with. Accordingly, two main ideas from our method are combined from [**?**] and [**?**]. First, we use a generator introduced in GAIN [**?**], which leverages a GAN to generate artificial missing values for data collection of different origins, including medical records. Second, likewise, to cGAN [**?**], which incorporates given auxiliary information into generative and discriminative functions, in our ccGAN formulation, we provide a significant step forward to our generator by conditioning it not only to the observable values but also to those fully-annotated features that are usually available in EHR datasets (e.g., age, weight, glycemia, etc.). Thus, our proposed ccGAN formulation provides novel insights within the ML and biomedical informatics community by leveraging the idea to combine fully

observed with partially observed predictors to improve the data imputation accuracy.

3. Multi-Center Diabetic dataset

The real clinical EHR Multi-Center Diabetic (MDC) dataset originally consisted of 120K diabetic patients and is structured in *demographics field* (i.e., patient's identification number (ID patient), gender, year of birth, diabetes diagnosis date); *pathological field* (i.e., ID patient, ICD-9 codes, pathology diagnosis date); and *lab tests field* (i.e., ID patient, lab tests codes, lab tests values, lab tests prescription date).

3.1. Definition of control and DR patients

The diabetologist selected all the ICD-9 codes of the *pathological field* associated with DR: the univocal ICD-9 code indicates a non-DR condition, while all the other ICD-9 codes indicate a DR condition.

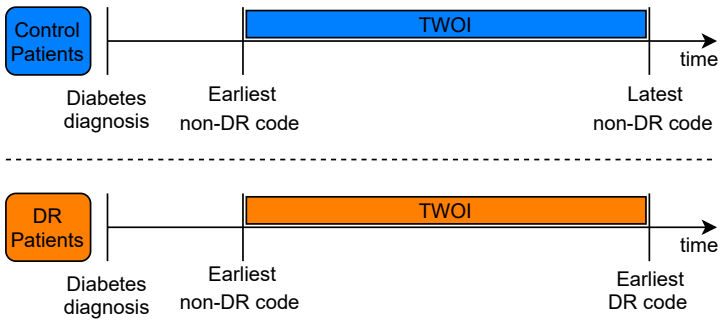

Figure 1: Observational time window of interest (TWOI) for control and DR patients.

All the ICD-9 codes that did not specify DR or non-DR conditions were removed from *pathological field*, bacause no longer of interest for the use case. Then, for every patient, both ICD-9 and lab test codes were removed if the pathology diagnosis date and lab tests prescription date preceded the diabetes diagnosis date (i.e., the use case temporal starting point). Figure 1 describes the inclusion criteria for selecting the time window of interest (TWOI) for control and DR patients.

3.1.1. Control patients - TWOI

A control patient was defined by at least two consecutive ICD-9 codes of non-DR and none of the DR codes within the TWOI. A TWOI of a control patient (see Figure 1 - upper side) is delimited by the earliest ICD-9 code of non-DR and the latest ICD-9 code of non-DR.

3.1.2. DR patients - TWOI

A DR patient was defined by at least an ICD-9 code of non-DR followed by one ICD-9 code of DR. A TWOI of a DR patient (see Figure 1 - bottom side) is delimited by the earliest ICD-9 code of non-DR and the earliest ICD-9 code of DR. A patient was included in the study only if the date of the earliest ICD-9 code of non-DR preceded the earliest date of ICD-9 code of DR.

3.2. Preprocessing

Following the definition of control and DR patients, the MDC dataset consists of 40555 patients (31611 control patients, 8944 DR patients) and 60 demographical and lab test features (predictors). The preprocessing procedure consists of feature analysis and patient selection stages.

3.2.1. Features analysis

A subset of 48 predictors was chosen by two diabetologists based on their experience in the clinical task of interest. Thus, the predictors were grouped by the distribution of their missing values (see Table 1). Predictors were split into green (X_g), yellow (X_y), and red (X_r) predictors according to the following criteria (see Figure 2):

- X_g contains less than 2% of missing values per patient and less than 50% of missing values for the whole dataset;
- X_y contains between 3% and 40% of missing values per patient and between 50% and 80% of missing values for the whole dataset;
- X_r contains more than 40% of missing values per patient and more than 80% of missing values for the whole dataset.

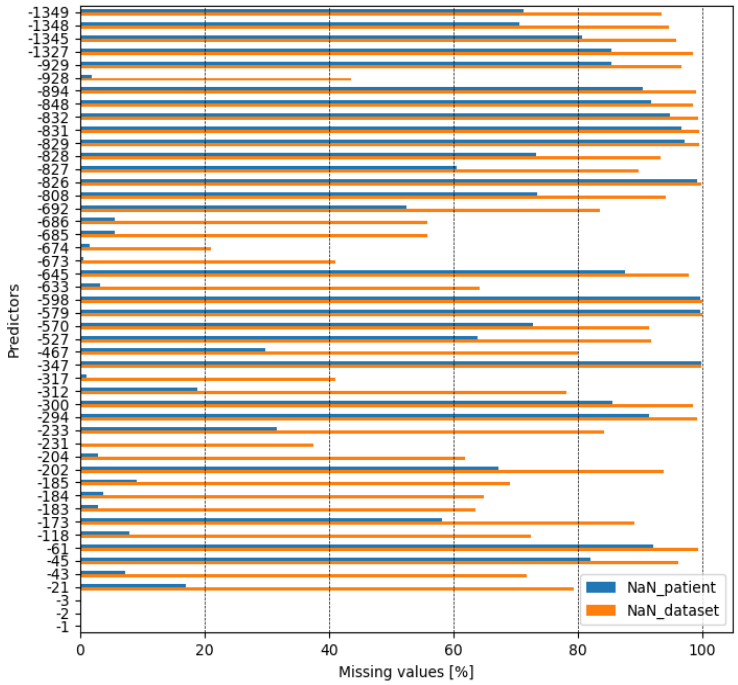

Figure 2: Missing values (NaNs) distribution over patients (blue) and over the whole MDC dataset (orange).

3.2.2. Patient selection

To obtain the X_g predictors filled (i.e., no missing values) across all the patients, we removed the 2981 patients (i.e., $\sim$ 80% control patients, 20% DR patients) that do not contain all the X_g predictors simultaneously. Table 2 describes the statistics of the MDC dataset after the patient selection preprocessing stage.

4. Method

The amount of observations for each patient in our real clinical MDC dataset is limited and sparsely distributed over time (see Table 2). Starting from this evidence, we justify our non-temporal data imputation approach by computing the auto-distance correlation, intending to verify whether there are no temporal dependencies within multiple observations of the same patient across time (see Section 4.1). Afterward, we present our ccGAN approach for data imputation on the MDC dataset.

Code	Description	Uom	Code	Description	Uom
-1349	Albumin to creatinine ratio	mg/mm	-570	Proteines (uri)	mg/dl
-1348	Creatinine clearance	ml/min	-527	Blood plates	1000/mm^3
-1345	Creatininuria	mg/dl	-467	Microalbuminuria	mg/l
-1327	Winsor index	Null	-347	Glicosuria	G/l
-929	Microalbuminuria	mg/24h	-317	Fasting glycaemia	mg/dl
-928	Body mass index	Kg/m^2	-312	Gamma-glutamyl transferase	UI/l
-894	Urine culture	Null	-300	Alkaline phosphatase	UI/l
-848	Potassium (uri)	mEq/l	-294	Fibrinogen (serum)	mg/dl
-832	Pre-prandial glycaemia	mg/dl	-233	Hemoglobin	g/dl
-831	Pre-dinner glycaemia	mg/dl	-231	Glycated hemoglobin	%
-829	Glycaemia h 23	mg/dl	-204	Creatinine	mg/dl
-828	Post-prandial glycaemia	mg/dl	-202	Creatine phosphokinase (serum)	UI/l
-827	Post-breakfast glycaemia	mg/dl	-185	LDL cholesterol	mg/dl
-826	Post-dinner glycaemia	mg/dl	-184	HDL cholesterol	mg/dl
-808	Creatinine clearance	ml/min	-183	Cholesterol (total)	mg/dl
-692	Urine ketones	mg/dl	-173	Weist	cm
-686	Diastolic pressure	mmHg	-118	Serum glutamic-oxaloacetic transaminase	UI/l
-685	Systolic pressure	mmHg	-61	Amylase	UI/l
-674	Height	cm	-45	Albumin excretion rate	mcg/min
-673	Weight	kg	-43	Alanine aminotransferase test	UI/l
-645	Urea	mg/dl	-21	Uric acid	mg/dl
-633	12-hour fasting triglycerides	mg/dl	-3	Gender	Null
-598	Sodium (uri)	mEq/l	-2	Age	years
-579	Albuminuria/creatinuria ratio	Null	-1	Diabetes duration	years

Table 1: Green predictors (X_g) indicate a very low presence of missing values, yellow predictors (X_y) indicate a mild presence of missing values, and red predictors (X_r) indicate a high presence of missing values according to the criteria defined in Section 3.2.1.

Description	Statistics
Total patients	37574
Control:	78%
DR:	22%
Gender	
Male:	56%
Female:	44%
Age (*years*)	68($\pm$12)
Diabetes duration (*years*)	12($\pm$8)
# of observations per patient	19($\pm$15)
Predictors	48
X_g:	8
X_y:	13
X_r:	27

Table 2: Statistics of the MDC dataset.

4.1. Auto-distance correlation function

Auto-distance correlation function (ADCF) measures temporal correlation across univariate time series [**?**]. The ADCF can be expressed as a V-statistic of order two, which is degenerate under the null hypothesis of independence. Thus, considering a traditional autocorrelation plot where the confidence intervals are got simultaneously

may be complex. Given this motivation, the $(1-\alpha)\%$ confidence intervals are computed simultaneously adopting the Monte Carlo simulation and the independent wild bootstrap approach [**?**]. We set the significance level $\alpha = 0.05$ and the number of bootstrap replications $b = 499$ to obtain the $(1-\alpha)\%$ empirical critical values. By exploring different lags (MaxLag=$5, 10, 15, 20, 25$) for computing the ACDF function within multiple observations of the same patient, we did not find any feature that overcame the critical value for more than the 5% of patients. Thus, the non-temporal correlation among the values of the predictors was evidenced.

Considering this finding, we employed a non-temporal configuration of the MDC dataset, where a single value for each predictor is considered for the i-th patient.

4.2. Clinical conditional Generative Adversarial Network (ccGAN)

In the standard cGAN formulation, two players minimax game between generative neural network G (generator) and discriminative neural network D (discriminator) is defined as follows:

$$\begin{aligned} min_G max_D \Big(& \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_{data}(\mathbf{x},\mathbf{y})}\big[\log D(\mathbf{x},\mathbf{y})\big] \\ & + \mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z}),\mathbf{y}\sim p_{data}(\mathbf{y})}\big[\log(1-D(G(\mathbf{z}|\mathbf{y}),y))\big]\Big). \end{aligned} \tag{1}$$

G and D are trained simultaneously, conditioned on some extra information $\mathbf{y}$. The generator learns a function that performs mapping for $\mathbf{z}$ from a simple distribution like $\mathscr{U}(0,1)$ to the distribution defined by data collection p_{data}. Thus, the generator aims to learn to produce samples indistinguishable from real data observations. On the contrary, the discriminator's objective is to separate generated samples from the real data accurately. Let $\mathbf{x}^g$, $\mathbf{x}^y$ and $\mathbf{x}^r$ are samples of green, yellow, and red predictors that are taking values in $\mathscr{X}^g = \mathscr{X}_1^g \times ... \times \mathscr{X}_{d_g}^g$, $\mathscr{X}^y = \mathscr{X}_1^y \times ... \times \mathscr{X}_{d_y}^y$ and $\mathscr{X}^r = \mathscr{X}_1^r \times ... \times \mathscr{X}_{d_r}^r$ spaces respectively. Distribution of the random variables $\mathbf{x}^g$, $\mathbf{x}^y$ and $\mathbf{x}^r$ are defined by corresponding data collections X_g, X_y, X_r and will denote $P(X_g)$, $P(X_y)$, $P(X_r)$. Considering the minimal amount of information provided by the matrix X_r (more than 40% of patients do not contain these features), we decided to impute only X_y predictors, conditioned on extra information given by X_g predictors. It is worth noting

that, differently from the state-of-the-art literature, the imputation of X_y still represents a challenging task, where the available information is highly limited and the proposed ccGAN approach should accurately impute between the 3% and 40% of missing values per patient. Accordingly, we consider a data collection $S = \{(\mathbf{x}_i^y, \mathbf{x}_i^g)\}_{i=1}^N = S_F \cup S_\perp$ of size N that consists of two subsets $S_F = \{(\mathbf{x}_i^y, \mathbf{x}_i^g) \in \mathscr{X}^y \times \mathscr{X}^g\}_{i=1}^{N_F}$ and $S_\perp = \{(\mathbf{x}_i^y, \mathbf{x}_i^g) \in \tilde{\mathscr{X}}^y \times \mathscr{X}^g\}_{i=1}^{N_\perp}$, where $\tilde{\mathscr{X}}^y = (\mathscr{X}_1^y \cup \{\perp\}) \times \ldots \times (\mathscr{X}_{d_y}^y \cup \{\perp\})$ and symbol $\perp$ indicates unobserved components ($N = N_F + N_\perp$ and $N_F \ll N_\perp$). Then further in our explanation, when referring to a sample $(\mathbf{x}^y, \mathbf{x}^g)$ drawn from $S_\perp$, we will use $(\tilde{\mathbf{x}}^y, \mathbf{x}^g)$.

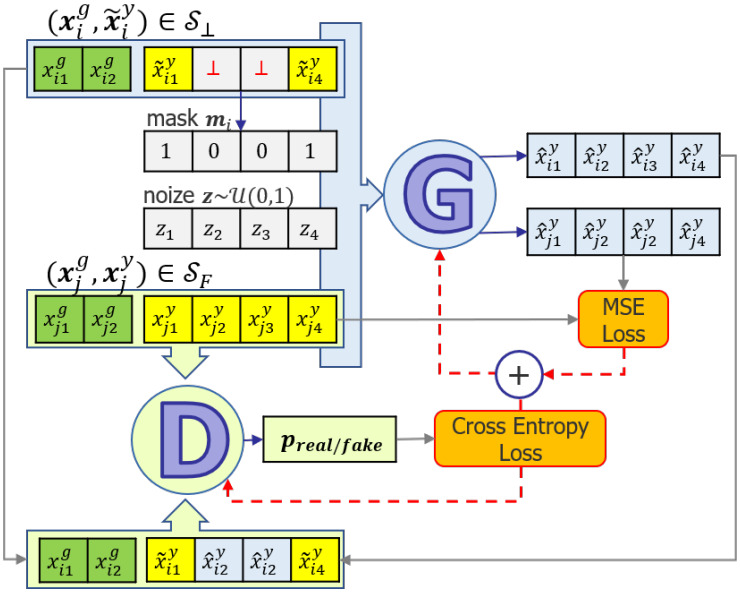

Figure 3: The proposed ccGAN architecture.

Architecture for our ccGAN-based imputation strategy is shown in Figure 3 and consists of two-players neural networks. First is a $G : \tilde{\mathscr{X}}^y \times \mathscr{X}^g \times \{0,1\}^{d_y} \to \mathscr{X}^y$ - generative neural network - that, conditionally on extra information given by green predictors $\mathbf{x}^g$ and partially available values of yellow predictors $\tilde{\mathbf{x}}^y$, performs mapping for the random variable $\mathbf{z}$ from distribution $\mathscr{U}(0,1)$ to corresponding complete vector $\mathbf{x}_{gen}^y$. Accordingly, if $\mathbf{m} \in \{0,1\}^{d_y}$ is a mask vector that indicates the availability of

each predictor value in $\tilde{\mathbf{x}}^{\mathbf{y}}$, then

$$\mathbf{x}^{y}_{gen} = G(\mathbf{m} \odot \tilde{\mathbf{x}}^{y}, \overline{\mathbf{m}} \odot \mathbf{z}, \mathbf{x}^{g}),$$

where $\overline{\mathbf{m}}$ denotes a complement of $\mathbf{m}$. Since the output of G consist of predictions of even non-missing values, the imputed vector is

$$\hat{\mathbf{x}}^{y} = \mathbf{m} \odot \tilde{\mathbf{x}}^{y} + \overline{\mathbf{m}} \odot \mathbf{x}^{y}_{gen}.$$

Next, similarly to cGAN, we define a discriminative neural network $D : \mathscr{X}^{y} \times \mathscr{X}^{g} \rightarrow [0,1]$ - an adversary to train G - which objective is to distinguish real full observations $(\mathbf{x}^{y}, \mathbf{x}^{g}) \in S_F$ from incomplete but imputed by G observations $(\hat{\mathbf{x}}^{y}, \mathbf{x}^{g})$, where $(\tilde{\mathbf{x}}^{y}, \mathbf{x}^{g}) \in S_{\perp}$. In particular, D and G are trained jointly so that D is optimized to maximize the probability of D predicting a correct label for a real or synthetic sample. In contrast, G is optimized to minimize the probability of D to identify generated samples. Then, discriminative loss for the minimax GAN optimization problem in the ccGAN model is the following:

$$\begin{aligned} L_d(G,D) = {} & \mathbb{E}_{(\mathbf{x}^{y}, \mathbf{x}^{g}) \in S_F} \left[\log D(\mathbf{x}^{y}, \mathbf{x}^{g}) \right] \\ & + \mathbb{E}_{(\tilde{\mathbf{x}}^{y}, \mathbf{x}^{g}) \in S_{\perp}, \mathbf{z}} \left[\log(1 - D(G(\tilde{\mathbf{x}}^{y}, \mathbf{x}^{g}, \mathbf{m}, \mathbf{z}), \mathbf{x}^{g}) \right]. \end{aligned} \tag{2}$$

Moreover, by taking into account that in our setup $N_F \neq 0$ and data is missing completely at random (MCAR), we use an additional term in the objective function - masked reconstruction loss - computed over real full samples to stabilize training of the introduced model. Specifically, let us define an operator f_{nan} that present missing values to the full vector of yellow predictors $\mathbf{x}^{y}$ from $(\mathbf{x}^{y}, \mathbf{x}^{g}) \in S_F$ with respect to the mask $\mathbf{m}$:

$$f_{nan}(\mathbf{x}^{y}, \mathbf{m}) = \mathbf{x}^{y} \odot \mathbf{m} + nan \odot \overline{\mathbf{m}}.$$

Accordingly, if $\mathbf{m}$ is sampled from the collection of masks in $S_{\perp}$, which is MCAR, then $(f_{nan}(\mathbf{x}^{y}, \mathbf{m}), \mathbf{x}^{g}) \in \tilde{\mathscr{X}^{y}} \times \mathscr{X}^{g}$, where $(\mathbf{x}^{y}, \mathbf{x}^{g}) \in S_F$; and masked reconstruction loss is

defined by:

$$L_r(G) = ||G(f_{nan}(\mathbf{x}^y, \mathbf{m}), \mathbf{x}^g, \mathbf{m}, \mathbf{z}) \odot \overline{\mathbf{m}} - \mathbf{x}^y \odot \overline{\mathbf{m}}||_2. \tag{3}$$

Finally, in the ccGAN imputation strategy, two players' minimax game between generator G and discriminator D is defined by two-part loss:

$$min_G max_D \Big(L_d(G,D) + L_r(G) \Big), \tag{4}$$

which we solve in a minibatch stochastic iterative manner described in Algorithm 1. Proposed method shares with the original GAN a property that global minimum is achieved if and only if $p_{data}(\mathbf{x}^y, \mathbf{x}^g) = p_g(\mathbf{x}^y_{gen}, \mathbf{x}^g)$, which can be proven as shown in [**?**].

Data: training set $S = S_F \cup S_\perp$
Initialization: $\theta_D^{(0)}, \theta_G^{(0)}$ # weights for G and D respectively
for $i = 0, \ldots, N_{epochs}$ **do**
  Draw minibatch $\mathscr{B}_F = \{\mathbf{x}^y_j, \mathbf{x}^g_j\}_{j=1}^{m_b}$ from S_F
  Draw minibatch $\mathscr{B}_\perp = \{\tilde{\mathbf{x}}^y_j, \mathbf{x}^g_j\}_{j=1}^{m_b}$ from $S_\perp$
  for $\mathscr{B}_\perp$ **do**
    $\mathbf{m}_j \leftarrow 1 - \mathbb{1}_\perp(\tilde{\mathbf{x}}^y_j)$
    $\hat{\mathbf{x}}^y_j \leftarrow G\big(\mathbf{m}_j \odot \tilde{\mathbf{x}}^y_j + \overline{\mathbf{m}}_j \odot \mathbf{z}_j, \mathbf{x}^g_j\big)$
    $\hat{\mathbf{x}}^y_j = \mathbf{m}_j \odot \tilde{\mathbf{x}}^y_j + \overline{\mathbf{m}}_j \odot \hat{\mathbf{x}}^y_j$
  end
  $L_D = \sum_{\mathscr{B}_F} \log D(\mathbf{x}^y_j, \mathbf{x}^g_j) + \sum_{\mathscr{B}_\perp} \log(1 - D(\hat{\mathbf{x}}^y_j, \mathbf{x}^g_j))$
  $L_G = \sum_{\mathscr{B}_\perp} \log(1 - D(\hat{\mathbf{x}}^y_j, \mathbf{x}^g_j))$
  for $\mathscr{B}_F$ **do**
    $\hat{\mathbf{x}}^y_j \leftarrow G\big(\mathbf{m}_j \odot \mathbf{x}^y_j + \overline{\mathbf{m}}_j \odot \mathbf{z}_j, \mathbf{x}^g_j\big)$
  end
  $L_G = L_G + \frac{1}{m_b} \sum_{\mathscr{B}_F} (\overline{\mathbf{m}}_j \odot \hat{\mathbf{x}}^y_j - \overline{\mathbf{m}}_j \odot \mathbf{x}^y_j)^2$
  $\theta_D^{(i+1)} \leftarrow Adam(-L_D, \theta_D^{(i)}, \alpha, \beta)$ # update of D
  $\theta_G^{(i+1)} \leftarrow Adam(L_G, \theta_G^{(i)}, \alpha, \beta)$ # update of G
end

Algorithm 1: Pseudo-code of ccGAN.[1]

X_g and the imputed X_y represent the predictors for each patient, while the label is represented in terms of control (0) and DR (1) patients.

[1] Extended code and data available upon request.

4.3. Experimental comparisons

We introduce the comparisons in terms of data imputation (see Section 4.3.1) and classification ML models (see Section 4.3.2) techniques.

4.3.1. Data imputation techniques

We start the experimental analysis by comparing the quality of imputed values predicted by the ccGAN method versus other state-of-the-art GAN-based missing data imputation strategies - baselines like GAIN and MisGAN. In this experiment, all the algorithms are trained with $S_{\perp}$ set together with a randomly selected subset of S_F. In their turn, the rest of the full observations form a set for testing. In other words, $S_F = S_F^{train} \cup S_F^{test}$, where we set train size proportion equal to 0.8. Then, after the model of choice g^* is trained, accuracy on the test set is evaluated by computing masked mean squared error (MSE) between estimated values of missing yellow predictors in a set $\{(f_{nan}(\mathbf{x}_i^y, \mathbf{m}), \mathbf{x}_i^g)\}_{(\mathbf{x}_i^y, \mathbf{x}_i^g) \in S_F^{test}}$ and their real values:

$$\frac{1}{\|S_F^{test}\|} \sum_{(\mathbf{x}_i^y, \mathbf{x}_i^g) \in S_F^{test}} \left(g^*(f_{nan}(\mathbf{x}^y, \mathbf{m}), \mathbf{x}^g) \odot \overline{\mathbf{m}} - \mathbf{x}^y \odot \overline{\mathbf{m}}\right)^2, \tag{5}$$

where m is sampled from the set of masks in $S_{\perp}$.

Moreover, once we evaluated the ccGAN in terms of data imputation performance with respect to the state-of-the-art GAN-based missing data imputation strategies, we compared our proposed ccGAN with other state-of-the-art data imputation techniques such as KNN [**?**], MissF [**?**], and MICE [**?**]. In this case, our goal was to compare the performance accuracy of the target label prediction in the case of training on complete data imputed by different imputation strategies. This experimental setup is explained next.

4.3.2. ML models, metrics, and experimental procedure

In the prediction stage, we used state-of-the-art ML models widely adopted for disease prediction [**?**], such as eXtreme Gradient Boosting (XGB), RF, Decision Tree (DT), Linear and Gaussian Support Vector Machine (SVM), Logistic Regression (LR), KNN. The predictive performance was evaluated by the following metrics: Accuracy,

macro-F1 (F1), macro-Precision (Precision), macro-recall (Recall), Area Under the receiver operating characteristic Curve (AUC), and Area Under the Precision-Recall Curve (PRAUC). We implemented a Tenfold Cross-Validation (CV-10) experimental procedure. The hyperparameters of the ML models were tuned in a nested Fivefold Cross-Validation by implementing a grid-search [**?**] and optimizing the PRAUC metric.

5. Results

In this section we first report both the imputation 5.1) and predictive (see Section 5.2) performance comparisons for the MDC dataset. Afterward, we provide the performed statistical analysis to highlight the possible significant predictive performance improvement of the proposed ccGAN approach with respect to state-of-the-art models (see Section 5.3). Finally, we report an additional experiment on a benchmark dataset (see Section 5.4) to evaluate the imputation and predictive performance comparisons for different missingness rates.

5.1. Imputation performance

According to the result of averaged masked MSE computed over 20 different random splits of S_F, ccGAN outperformed the baseline models GAIN and MisGAN (see Table 3). GAIN was the best competitor.

Imputation Model	**masked MSE**
GAIN	0.192 ± 0.018
MisGAN (the imputer)	0.203 ± 0.015
ccGAN	$\mathbf{0.154 \pm 0.015}$

Table 3: Imputation performance in terms of masked MSE of different GAN-based models for missing data imputation and proposed ccGAN model averaged over 20 random training/test data splits.

Next, once proven that ccGAN outperformed baseline GAN-based missing data imputation strategies, we present the results of the proposed ccGAN in terms of predictive performance.

5.2. Predictive performance

The XGB and RF performed best among other tested supervised classifiers, such as DT, SVM, LR, and KNN. Thus, we show only the predictive performance of the proposed data imputation approach by applying XGB (see Table 4) and RF (see Table 5) as ML classification models. It is worth noting that we exploited the proposed ccGAN for imputing the value of X_y predictors. The overall best predictive performance was reached by the RF model in X_g+X_y setting adopting our proposed ccGAN imputation technique (PRAUC = 66.16 ± 1.09). The best imputation technique competitor is represented by MICE (PRAUC = 65.53 ± 1.04). Employing single X_y or X_g predictors leads to decreased performance. The same trend was reached by the XGB model but with globally lower predictive performance than the RF model. However, the ccGAN imputation technique in X_g+X_y setting (PRAUC = 65.20 ± 1.09) remains the best strategy.

Predictors	Accuracy	F1	Precision	Recall	AUC	PRAUC
X_y (KNN)	83.12 ± 0.39	66.95 ± 0.86	80.51 ± 1.18	64.06 ± 0.67	76.09 ± 0.68	58.16 ± 1.22
X_y (missF)	82.88 ± 0.35	66.90 ± 0.73	79.14 ± 1.06	64.10 ± 0.57	76.27 ± 0.76	58.44 ± 0.87
X_y (MICE)	82.79 ± 0.38	67.20 ± 0.91	78.25 ± 0.93	64.43 ± 0.74	77.53 ± 0.73	59.41 ± 1.40
X_y **(ccGAN)**	83.06 ± 0.47	67.85 ± 1.07	78.89 ± 1.21	64.96 ± 0.88	77.89 ± 0.55	60.25 ± 1.22
X_g+X_y (KNN)	83.66 ± 0.42	69.18 ± 0.86	80.36 ± 1.21	66.04 ± 0.71	79.69 ± 0.69	62.85 ± 1.06
X_g+X_y (missF)	83.78 ± 0.41	69.59 ± 0.83	80.41 ± 1.08	66.42 ± 0.68	80.08 ± 0.70	63.35 ± 1.23
X_g+X_y (MICE)	83.81 ± 0.55	69.89 ± 0.98	80.16 ± 1.54	66.73 ± 0.78	80.97 ± 0.50	64.15 ± 1.39
X_g+X_y **(ccGAN)**	84.12 ± 0.45	70.68 ± 0.75	80.67 ± 1.29	67.43 ± 0.59	81.40 ± 0.45	$\mathbf{65.20 \pm 1.09}$
X_g	82.67 ± 0.49	67.90 ± 0.80	76.95 ± 1.45	65.20 ± 0.63	74.70 ± 0.80	57.28 ± 1.09

Table 4: Predictive performance of XGB model in X_g, X_y, and $X_g + X_y$ settings: comparison between our proposed data imputation techniques and other competitors. The best predictive performance result in terms of PRAUC is reported in bold.

5.3. Statistical analysis

According to the Anderson-Darling test, the PRAUC scores in the CV-10 experimental procedure deviated from normality ($p < .01$). Hence, the statistical comparison between our proposed ccGAN approach and the best data imputation competitors was performed through a non-parametric, one-sided Wilcoxon signed-rank test ($\alpha = 0.05$) for the RF and XGB models. The performance of the ccGAN ($X_g + X_y$ setting) was significantly greater ($p < 0.05$) than MICE ($X_g + X_y$ setting) by applying the RF model.

Predictors	Accuracy	F1	Precision	Recall	AUC	PRAUC
X_y (KNN)	83.84 ± 0.32	65.70 ± 0.89	89.55 ± 0.86	62.80 ± 0.65	78.12 ± 0.71	59.03 ± 1.09
X_y (MissF)	83.74 ± 0.31	65.80 ± 0.93	88.77 ± 0.68	62.90 ± 0.68	78.76 ± 0.57	60.00 ± 1.21
X_y (MICE)	83.74 ± 0.34	66.30 ± 0.86	87.20 ± 1.20	63.33 ± 0.63	79.54 ± 0.69	60.84 ± 1.21
X_y **(ccGAN)**	83.70 ± 0.44	66.00 ± 1.17	86.60 ± 1.23	63.10 ± 0.86	80.00 ± 0.78	61.60 ± 1.31
X_g+X_y (KNN)	84.23 ± 0.37	68.10 ± 0.84	86.43 ± 1.35	64.75 ± 0.64	81.10 ± 0.82	64.16 ± 1.23
X_g+X_y (MissF)	84.28 ± 0.27	68.21 ± 0.59	86.64 ± 1.06	64.81 ± 0.45	81.38 ± 0.76	64.65 ± 1.03
X_g+X_y (MICE)	84.30 ± 0.39	68.36 ± 0.86	86.47 ± 1.36	64.94 ± 0.65	82.28 ± 0.50	65.53 ± 1.04
X_g+X_y **(ccGAN)**	84.30 ± 0.45	68.60 ± 1.05	86.10 ± 1.29	65.14 ± 0.81	82.67 ± 0.58	$\mathbf{66.16 \pm 1.09}$
X_g	83.91 ± 0.28	67.89 ± 0.70	84.28 ± 0.77	64.66 ± 0.55	78.12 ± 0.54	60.50 ± 1.05

Table 5: Predictive performance of RF model in X_g, X_y, and $X_g + X_y$ settings: comparison between our proposed data imputation techniques and competitors. The best predictive performance result in terms of PRAUC is reported in bold.

Furthermore, the performance of the ccGAN ($X_g + X_y$ setting) continued to be significantly greater ($p < 0.05$) than MICE ($X_g + X_y$ setting) by applying the XGB model.

5.4. Additional experiment on a benchmark dataset: MIMIC-III

To evaluate the robustness of our proposed ccGAN approach across different missingness rates [**?**], we adopt the benchmark MIMIC-III dataset [**? ?**]. Unlike the real MDC dataset, where the missingness is naturalistic, we artificially generate missing values in the MIMIC-III dataset to purposefully accomplish this aim. We describe the MIMIC-III dataset and the synthetic missing values generation in Section 5.4.1. We describe the results in terms of imputation and predictive performance in Section 5.4.2 and Section 5.4.3, respectively.

5.4.1. Benchmark MIMIC-III dataset

We extracted a subset (i.e., Feature set A [**?**]) from the original MIMIC-III dataset and then averaged the feature observations over time. The extracted feature set consists of 17 features utilized in calculating the SAPS-II score. To evaluate the robustness of the ccGAN imputation performance according to different missingness rates, among all the patients, we selected only the 8509 patients with fully observable features (i.e., no missing values)[**?**]. The features such as age, gender, admission type, and Glasgow coma scale were selected as X_g. In contrast, for the remaining ones (X_y) different missingness rates were artificially generated as follows: we all the time chosen a proportion of patients (frac_mis $= [0.3, 0.6, 0.9]$) whose we developed missing values

by using a random mask with a binomial distribution with different probabilities (p $= [0.3, 0.6, 0.9]$). The binary classification task consisted of in-hospitality mortality prediction. The class imbalance was almost similar to the MDC dataset (see Table 2): the 20% of the patients died during the hospital stay after being admitted to an ICU, the remaining ones remained alive.

5.4.2. Imputation performance

For the MIMIC-III dataset, according to the imputation performance of averaged masked MSE computed over 20 different random splits, ccGAN outperformed the baseline GAIN and MisGAN models (see Table 6) when the missingness rate increases (i.e., frac_mis $= 0.6 - 0.9$ and p $= 0.6 - 0.9$).

frac_mis	p	**ccGAN**	GAIN	MisGAN
0.3	0.3	$\mathbf{0.316 \pm 0.024}$	0.330 ± 0.039	0.484 ± 0.052
	0.6	0.646 ± 0.036	$\mathbf{0.636 \pm 0.058}$	0.979 ± 0.052
	0.9	$\mathbf{1.010 \pm 0.044}$	1.414 ± 0.383	1.651 ± 0.121
0.6	0.3	0.324 ± 0.047	$\mathbf{0.320 \pm 0.050}$	0.442 ± 0.023
	0.6	$\mathbf{0.618 \pm 0.025}$	0.664 ± 0.142	0.972 ± 0.083
	0.9	$\mathbf{1.004 \pm 0.047}$	1.115 ± 0.137	1.618 ± 0.128
0.9	0.3	0.388 ± 0.158	$\mathbf{0.308 \pm 0.042}$	0.458 ± 0.067
	0.6	$\mathbf{0.613 \pm 0.038}$	0.677 ± 0.109	0.970 ± 0.153
	0.9	$\mathbf{1.028 \pm 0.055}$	2.097 ± 2.451	1.618 ± 0.157

Table 6: Imputation performance for the MIMIC-III dataset in terms of masked MSE of different GAN-based models for missing data imputation and proposed ccGAN model averaged over 20 random training/test data splits. Respectively, frac_mis $= [0.3, 0.6, 0.9]$ and p $= [0.3, 0.6, 0.9]$ represent the fraction of patients whose we generated missing values and the probability of the binomial distribution.

5.4.3. Predictive performance

Even in the MIMIC-III dataset, the XGB and RF models achieved the best predictive performance results among other tested supervised classifiers, such as DT, SVM, LR, and KNN. Thus we show only the XGB (see Figure 4) and RF (see Figure 5) predictive performance results in terms of PRAUC for the X_g+X_y setting across different missingness rates.

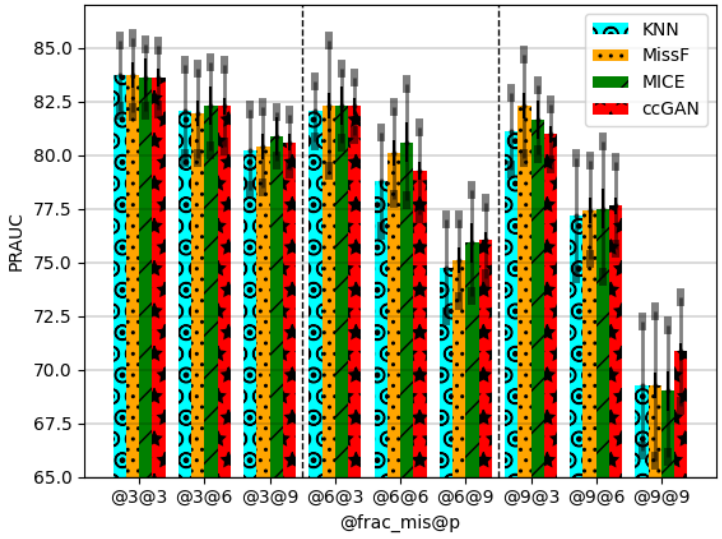

Figure 4: The XGB predictive performance in the MIMIC-III dataset regarding PRAUC. KNN, MissF, MICE, and ccGAN imputation techniques were compared across different missingness rates. Respectively, frac_mis $= [0.3, 0.6, 0.9]$ and p $= [0.3, 0.6, 0.9]$ represent the fraction of patients whose we generated missing values and the probability of the random mask binomial distribution.

The XGB model associated with our proposed ccGAN imputation technique outperformed the other competitors in the most challenging settings (i.e. high missingness rates): frac_mis $= 0.6$, p $= 0.9$; frac_mis $= 0.9$, p $= 0.6$; and frac_mis $= 0.9$, p $= 0.9$. Also, the RF model associated with our proposed ccGAN imputation technique outperformed the other competitors in the most challenging settings (i.e. high missingness rates): frac_mis $= 0.3$, p $= 0.9$; frac_mis $= 0.6$, p $= 0.9$; and frac_mis $= 0.9$, p $= 0.9$. Similarly to the imputation performance results (see Section 5.4.2), the higher the p and frac_mis, the more ccGAN performs better for both XGB and RF. In fact, for the frac_mis $= 0.9$, p $= 0.9$ setting, ccGAN achieved the highest PRAUC gain with respect to KNN (1.61% and 1.17%), MissF (1.61% and 2.07%) and MICE (1.85% and 1.61%) for both XGB and RF, respectively.

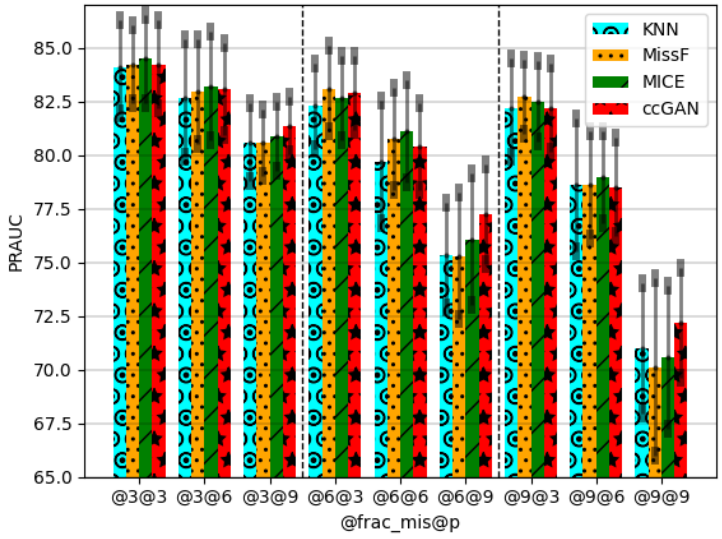

Figure 5: The RF predictive performance in the MIMIC-III dataset regarding PRAUC. KNN, MissF, MICE, and ccGAN imputation techniques were compared across different missingness rates. Respectively, frac_mis $= [0.3, 0.6, 0.9]$ and p $= [0.3, 0.6, 0.9]$ represent the fraction of patients whose we generated missing values and the probability of the random mask binomial distribution.

6. Discussion

For the proposed ccGAN approach, we discuss the impact on the MDC dataset (see Section 6.1) and its robustness across different missingness rates on the benchmark MIMIC-III dataset (see Section 6.2).

6.1. Impact on the Multi Diabetic Centre dataset

The imputation performance comparisons highlighted how the proposed ccGAN strategy was a reliable solution with respect to baseline GAN-based strategies for imputing missing values in the clinical MDC dataset under the MCAR assumption. Moreover, the predictive performance results and the statistical analysis showed that our proposed ccGAN strategy overcame the other state-of-the-art data imputation approaches for all ML models and all different settings. Our proposed ccGAN data imputation strategy was robust and effective in dealing with challenging real EHR datasets characterized by high sparsity, imbalanced settings, and noisy and redundant features. Even

if the X_g+X_y setting achieved the best predictive performance, the experimental results showed that the single X_y set contains more discriminative power than the single X_g set. This fact motivates that a correct and ad-hoc missing values imputation mechanism could be crucial to obtain a satisfactory predictive performance on routine EHR data.

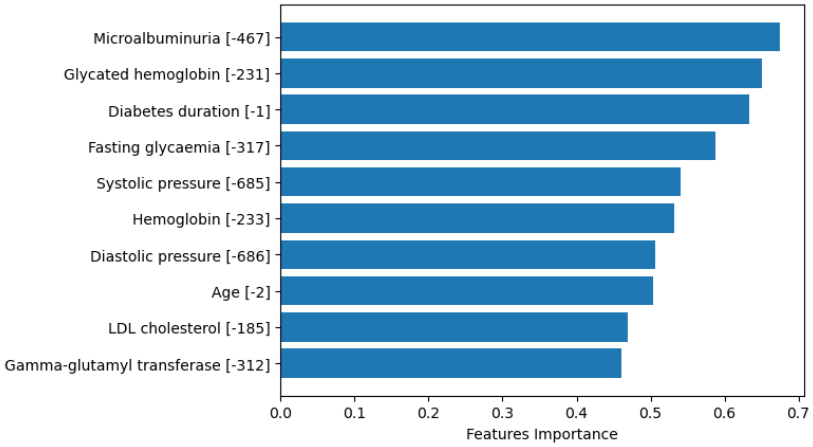

Figure 6: Top-10 discriminative predictors: X_g+X_y setting with ccGAN imputation strategy and RF model.

This assumption was also supported by the posterior interpretability analysis of the employed ML models. We computed the feature importance as the impurity accumulation decreased within each tree. The impurity decrease was computed for each feature and was averaged over each CV-10 run (see Figure 6). The microalbuminuria $\in X_y$ was found to be the most important predictor in the X_g+X_y setting using the ccGAN strategy associated with the RF model. This outcome highlights how the proposed data imputation strategy consistently imputes missing information, representing a discriminative pattern left unseen if we had used the original missing data as predictors. Besides the predictive performance output, this additional information assumes a crucial aspect in the clinical decision-making process.

A limitation of this work might be the exclusion of the X_r predictors (i.e., more than 40% of missing values per patient and more than 80% of missing values for the whole dataset) in the data imputation mechanism. These features, also if selected as necessary

by the diabetologists, were not imputed due to too high missingness rates. Another critical limitation might be excluding other EHR fields, such as exams and drug prescriptions. These fields may contain a considerable amount of missing information related to the availability of a generic drug code (parent code) and the missing of a specific unique drug code (child code). In future work, we aim to exploit a multi-view learning strategy that encapsulates our ccGAN data imputation approach for imputing missing values conditioned to different (eventually missing) views of the MDC dataset. The ccGAN data imputation strategy will also be extended to other diabetes complications by developing a fully-equipped clinical platform for managing diabetic patients.

6.2. Robustness for different missingness rates on the benchmark MIMIC-III dataset

The imputation and predictive performance results achieved from the benchmark MIMIC-III dataset justified the employment of the proposed ccGAN approach in a setting closer to the real clinical setting. In that scenario, the lab test exams are sparsely prescribed across patients; thus, the missingness rate is significantly high. Indeed, the additional experiment conducted on a benchmark MIMIC-III dataset proved how the ccGAN approach was the best to impute a significant number of missing values (90% of the patients have a 90% probability of containing missing values). Additionally, the ccGAN approach can be easily extended to other types of tasks besides diabetes complications-oriented predictions and this characteristic suggests its potential scalability in several predictive medicine tasks.

7. Conclusions

We proposed a novel ccGAN architecture capable of imputing missing values from routine EHR data collected from a multi-diabetic centers platform. We demonstrated how the proposed data imputation strategy was consistent for predicting DR in high missingness rates (i.e., between 3% and 40% of patients have the candidate feature missing). Within a DR screening program, our method is currently integrated into a clinical decision support system. It permits discovering the most discriminative predictors by also considering the missing information. Finally, we also demonstrated the

generalizability and robustness of the proposed ccGAN approach to solving a benchmark precision medicine task in more challenging missing values conditions.

Acknowledgements

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. This work was supported by the collaboration between Università Politecnica delle Marche, University of Macerata, METEDA srl, and AI Medical srl.

Ethics Statement The Ethical Committees of the University approved the experimental study and its guidelines as a clinical non-interventional (observational) study. EHR data are anonymous and their use, detention and conservation are regulated by an agreement between the University and data owners. All the process is inside the EU GDPR regulation. The proposed ccGAN approach is also compliant with the ethics guidelines of the European Commission (Human Agency and Oversight [**?**]). and it is currently integrated into a Clinical Decision Support System for screening purposes.