Journal of Computational Science 99 (2026) 102940

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

Assessing the impact of synthetic data generated by Bayesian networks on heart disease prediction☆

Ilaria Lazzaro [a],*, Marianna Milano [b], Allan Tucker [c], Mario Cannataro [a]

[a] Data Analytics Research Center, University "Magna Græcia" of Catanzaro, Department of Medical and Surgical Sciences, University "Magna Græcia" of Catanzaro, Viale Europa, Catanzaro, 88100, Italy
[b] Department of Experimental and Clinical Medicine, University "Magna Græcia" of Catanzaro, Viale Europa, Catanzaro, 88100, Italy
[c] Computer Science Department, Brunel University of London, Kingstone lane, Uxbridge UB, UB8 3PH, UK

ARTICLE INFO

Keywords:
Bayesian networks
Synthetic data generation
Heart disease prediction
Data quality

ABSTRACT

Synthetic data generation using Bayesian networks (BN) offers a promising approach to overcoming data scarcity in clinical prediction tasks, yet its actual impact on model performance remains underexplored. This study investigates the use of Bayesian network-based generative models to produce synthetic patient data and examines how the quality of the original real data influences the effectiveness of such augmentation. Three benchmark datasets from the UCI Heart Disease repository (Cleveland, Hungary, and Switzerland) were employed, all sharing an identical structure comprising 13 clinical predictors. The Cleveland dataset, which is the most complete and consistent among the three, was used exclusively as the training source for learning the Bayesian network structure and parameters under clinically informed constraints. To ensure robust evaluation, the dataset was partitioned into two independent subsets: 153 patients were used to train the Bayesian network, while 150 held-out patients were used exclusively to generate synthetic records. Predictive models were trained under three configurations: real data only, synthetic data only, and a hybrid real + synthetic (filtered) dataset, and evaluated using 10-fold cross-validation and external validation on independent cohorts. Results indicate that integrating real and synthetic data significantly improved accuracy and precision, particularly for the Switzerland cohort ($F(2,27)=23.06$, $\eta^2=0.63$), whereas improvements were smaller and partially non-significant in the noisier Hungarian dataset. These findings demonstrate that the effectiveness of synthetic augmentation depends on the structure and completeness of the source data, underscoring the importance of data quality for reliable generative modelling in clinical prediction.

1. Introduction

Over the past two decades, data-driven approaches have rapidly transformed healthcare analytics, particularly with the rise of electronic health records and scalable machine learning systems [1,2]. Beam and Kohane [1] emphasise how these methods enable population-level prediction, while [2] highlight their role in integrated digital medicine. This transformation has been particularly evident in cardiovascular disease prediction, where machine learning models show considerable promise for early diagnosis, risk stratification, and clinical decision support, as heart disease remains one of the leading causes of morbidity and mortality worldwide [3]. However, the effectiveness of these models often depends critically on the quantity, quality, and representativeness of available data.

While machine learning techniques excel at identifying complex, non-linear relationships within clinical datasets, their performance remains fundamentally constrained by the underlying data characteristics. Medical datasets present several challenges that limit model development and deployment. Many clinical studies draw from single institutions or relatively small patient populations, constraining both statistical power and the ability to develop generalisable models. Additionally, missing values and inconsistent records further degrade data quality, requiring data preprocessing and imputation. These challenges often result in predictive models that are prone to overfitting, lack robustness, or fail to generalise beyond the training environment.

To address these fundamental constraints, synthetic data generation has gained attention as a complementary strategy for dataset augmentation [4]. When properly implemented, synthetic data can help address

☆ This article is part of a Special issue entitled: 'AI&CS4Social' published in Journal of Computational Science.

* Corresponding author.
E-mail addresses: ilaria.lazzaro@unicz.it (I. Lazzaro), m.milano@unicz.it (M. Milano), Allan.Tucker@brunel.ac.uk (A. Tucker), cannataro@unicz.it (M. Cannataro).

https://doi.org/10.1016/j.jocs.2026.102940
Received 19 February 2026; Received in revised form 5 May 2026; Accepted 10 June 2026
Available online 15 June 2026

class imbalance, expand dataset diversity, and provide additional training examples while preserving patient privacy, a particularly important consideration in healthcare applications.

Bayesian networks (BN) offer a compelling framework for synthetic data generation in medical contexts [5]. Moreover, their structure can incorporate existing clinical knowledge, ensuring that generated samples reflect established medical understanding rather than purely empirical patterns.

BN provide a principled framework for modelling probabilistic dependencies in healthcare data [5]. Unlike purely data-driven models such as Generative Adversarial Network (GAN) or Variational Autoencoder (VAE), they can embed prior clinical knowledge directly into the network structure, preserving interpretability and causal coherence—an aspect emphasised by Lappenschaar et al. [5] and later reviewed by Pezoulas et al. [6]. In this regard, Kaur et al. [7] applied Bayesian networks to generate synthetic health records, focusing on the preservation of rare events and evaluating the extent to which synthetic datasets could maintain key statistical properties of the original data. Unlike their work, which primarily emphasised data fidelity and privacy considerations, this work explores the use of Bayesian network–based synthetic data generation specifically as a strategy aimed at improving predictive performance in cardiovascular disease modelling. This combination of interpretability, domain knowledge integration, and principled uncertainty quantification makes Bayesian networks particularly well-suited for healthcare applications where model transparency and clinical plausibility are essential, as highlighted in previous studies [5].

Three main research objectives guide our study:

- to determine whether models trained with synthetic augmentation outperform those relying solely on limited real-world datasets;
- to investigate how the intrinsic quality of real data influences the fidelity and predictive value of synthetic records;
- to analyse the impact of random sampling variability on the characteristics of generated datasets and downstream model performance.

By systematically addressing these questions, our study provides new insights into the practical value of Bayesian networks for healthcare data augmentation [8]. Ultimately, this work aims to clarify the role of synthetic data in overcoming common limitations of medical datasets and to inform the development of more reliable and generalisable predictive models in clinical practice. The remainder of this paper is structured as follows. Section 2 outlines the study objectives, Section 3 details the analysis workflow and methodologies employed, Section 4 presents the results of the analysis, and Section 5 concludes the paper underlying potential directions for future research.

2. Study objectives

In this study, we investigate the impact of synthetic data generated through Bayesian networks on predictive modelling in the medical domain, with a particular focus on heart disease datasets [9]. The underlying hypothesis is that, when carefully constructed, synthetic data can compensate for the limited size, class imbalance, and variability commonly encountered in real-world clinical datasets, thereby enhancing model performance without compromising data integrity or generalisability.

To systematically evaluate this hypothesis, we define three key research questions (RQ):

1. **RQ1:** Do synthetic data generated via Bayesian networks improve predictive performance compared to models trained exclusively on real data in small heart disease datasets?
   This question aims to determine whether synthetic augmentation offers tangible benefits in terms of accuracy, recall, or F1-score when real data are scarce.
2. **RQ2:** How does the initial quality of real data influence the effectiveness of synthetic data?
   We hypothesise that well-structured, informative real data support the generation of useful synthetic records, while noisy or unbalanced data may limit their utility or introduce bias.
3. **RQ3:** What is the impact of randomness in the sampling of real data on the characteristics of the synthetic data and the resulting model performance?
   To explore this, we train multiple Bayesian networks on different random subsets of the real dataset, examining the variability in both the generated data and model outcomes.

3. Materials and methods

Building upon the theoretical foundations of Bayesian networks and their application to probabilistic modelling, this section outlines the experimental methodology adopted in the present study. The entire workflow, from data selection to model evaluation, is illustrated in Fig. 1 and is described in detail in the following subsections.

3.1. Bayesian Networks: Background and construction from data

Bayesian Networks (BNs) are probabilistic graphical models that combine graph theory and Bayesian inference to represent uncertainty and conditional dependencies among random variables. A BN is defined by a directed acyclic graph (DAG), in which nodes represent variables and edges encode direct probabilistic dependencies, together with a set of local conditional probability distributions associated with each node.

From a data-driven perspective, the construction of a Bayesian network typically follows a structured pipeline. First, raw data are preprocessed to handle missing values and, when necessary, continuous variables are discretised to improve model stability and interpretability. Second, the network structure is learned either purely from data or under the guidance of prior domain knowledge, which can be expressed through structural constraints. Third, once the graph topology is defined, the parameters of the local conditional distributions are estimated from the data. Finally, the resulting model can be used for probabilistic inference, enabling reasoning under uncertainty and supporting classification or prediction tasks.

The mathematical foundation of Bayesian networks lies in Bayesian inference, which formalises how prior beliefs are updated in the presence of observed evidence. This framework allows the integration of expert knowledge and empirical data within a unified probabilistic representation. The conditional independence assumptions encoded in the DAG lead to a factorisation of the joint probability distribution, significantly reducing model complexity while preserving interpretability.

Based on this general construction process, the following sections describe how these steps were applied in the present study, starting with a description of the dataset 3.2, moving on to the definition of the network structure 3.3, 3.4 based on clinically informed constraints, and finally proceeding to parameter estimation and probabilistic inference 3.5.

3.2. Datasets

The experimental framework employs three benchmark datasets from the UCI Heart Disease repository, representing patient records from distinct geographic sites: Cleveland, Hungary, and Switzerland. All three datasets share the same structure and variable definitions, enabling consistent modelling and comparative evaluation across cohorts.

Each record is described by 13 clinical attributes capturing demographic information, symptoms, laboratory measurements, and exercise-related indicators, together with a target variable (`HeartDisease`) indicating the presence (values 1–4) or absence (0) of coronary artery disease.

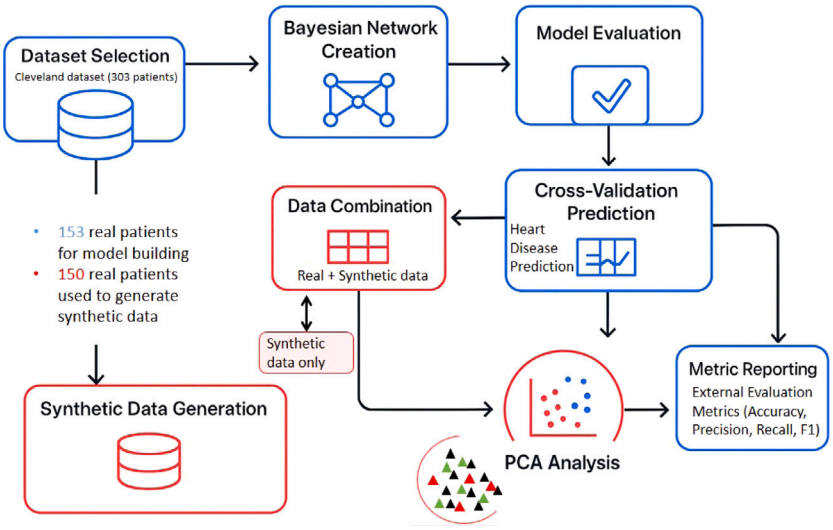

Fig. 1. Overview of the experimental workflow. The pipeline begins with dataset selection, followed by Bayesian Network construction and the generation of synthetic patient data. Real and synthetic samples are subsequently integrated and subjected to Principal Component Analysis (PCA) to evaluate distributional overlap and variability. Predictive models are trained and validated using cross-validation protocols, and their performance is quantified using standard classification metrics, including external validation on independent datasets.

The clinical attributes are encoded using standard abbreviations defined in the original dataset specification, including `age` (age in years), `sex` (biological sex), `cp` (chest pain type), `trestbps` (resting blood pressure), `chol` (serum cholesterol), `fbs` (fasting blood sugar), `restecg` (resting electrocardiographic results), `thalach` (maximum heart rate achieved), `exang` (exercise-induced angina), `oldpeak` (ST depression induced by exercise), `slope` (slope of the peak exercise ST segment), `ca` (number of major vessels coloured by fluoroscopy), and `thal` (thalassemia).

A detailed overview of data completeness and missing values for each variable across the three datasets is reported in Tables 1 and 2.

3.2.1. Cleveland dataset

This dataset comprises 303 patient records collected between May 1981 and September 1984 at the Cleveland Clinic Foundation (USA). The Cleveland dataset has relatively few missing values: 6 records contain incomplete information, primarily affecting the variables `ca` and `thal`. These records were retained after applying mode imputation to preserve as much original data as possible.

Due to its completeness and balanced class distribution, the Cleveland dataset was selected as the primary source for both training the Bayesian network and generating synthetic data [10].

3.2.2. Hungarian dataset

The Hungarian dataset contains 294 patient records collected at the Hungarian Institute of Cardiology in Budapest. It follows the same structure as the Cleveland dataset, with 13 clinical attributes and the `Heart disease` target variable. However, it exhibits a high proportion of missing values. Most records are incomplete, particularly in the variables `ca` (291 missing), `thal` (266), and `slope` (190), along with additional gaps in `chol`, `fbs`, and others.

Out of 294 records, only 1 is fully complete, while the remaining 293 have at least one missing value. This dataset was used exclusively for external validation to assess model generalisation [11].

3.2.3. Switzerland dataset

The Switzerland dataset consists of 123 patient records collected at the University Hospital in Zurich. As with the other datasets, it includes the standard 13 clinical features and the `Heart disease` target variable. This dataset is notably sparse, with missing data in critical variables such as `ca` (118 missing), `thal` (52), and `fbs` (75). Additional missing values appear in variables including `slope`, `oldpeak`, and `trestbps`. No records are fully complete: 100% of entries contain at least one missing value. Despite this limitation, the Switzerland dataset serves as an external test set to challenge the robustness and adaptability of the model [12].

A detailed overview of data completeness for each variable across the three datasets is presented in Tables 1 and 2.

Table 1
Number of complete values per variable across the Cleveland, Hungarian, and Switzerland datasets.

Variable	Cleveland	Hungarian	Switzerland
Age	303	294	123
Sex	303	294	123
Chest pain (cp)	303	294	123
Resting BP (trestbps)	303	293	121
Cholesterol (chol)	303	271	123
Fasting blood sugar (fbs)	303	286	48
Resting ECG (restecg)	303	293	122
Max heart rate (thalach)	303	293	122
Exercise angina (exang)	303	293	122
ST depression (oldpeak)	303	294	117
ST slope (slope)	303	104	106
Number of vessels (ca)	299	3	5
Thalassemia (thal)	301	28	71

Table 2
Number of missing values per variable across the Cleveland, Hungarian, and Switzerland datasets.

Variable	Cleveland	Hungarian	Switzerland
Age	0	0	0
Sex	0	0	0
Chest pain (cp)	0	0	0
Resting BP (trestbps)	0	1	2
Cholesterol (chol)	0	23	0
Fasting blood sugar (fbs)	0	8	75
Resting ECG (restecg)	0	1	1
Max heart rate (thalach)	0	1	1
Exercise angina (exang)	0	1	1
ST depression (oldpeak)	0	0	6
ST slope (slope)	0	190	17
Number of vessels (ca)	4	291	118
Thalassemia (thal)	2	266	52

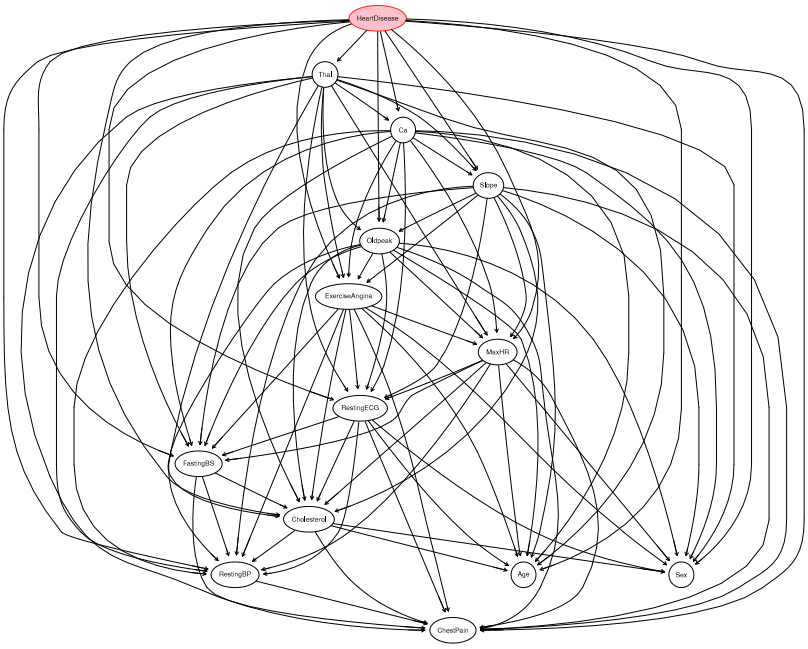

Fig. 2. Structure of the Bayesian network learned from the Cleveland dataset.

3.2.4. Data preprocessing

To enable structure learning via Bayesian networks, all continuous variables in the Cleveland dataset were discretised into three ordinal intervals using clinically relevant thresholds (see e.g. ACC/AHA guidelines for BP and standard clinical cut-offs for total cholesterol) [13, 14].

Specifically: `Age` was grouped into ≤45, 46–60, and >60 years; `RestingBP` into ≤120, 121–140, and >140 mmHg; `Cholesterol` into ≤200, 201–250, and >250 mg/dL; `MaxHR` into ≤110, 111–150, and >150 bpm; and `Oldpeak` into ≤1.0, 1.1–2.5, and >2.5. This binning strategy enhances interpretability and improves modelling stability in probabilistic graphical models [15–17].

Missing values were imputed using mode substitution, applied independently to each variable to preserve distributional consistency and minimise information loss.

3.3. Bayesian network modelling

In this study, we propose a network pipeline derived from clinical data using a constraint-based structural learning approach, which is subsequently employed for classification and inference tasks related to heart disease. The overall model, shown in Fig. 2, was constructed from a subset of the Cleveland dataset and encodes the joint probability distribution over the observed features and the target outcome.

The rest of this section describes the methodology applied for structured network learning. Specifically, we outline the procedure for structure learning based on clinical constraints, followed by parameter estimation of the conditional distributions, and conclude with the application of the trained network for probabilistic inference and classification.

3.4. Bayesian network structure learning

To avoid information leakage and ensure a clear separation between model training and synthetic data generation, the Cleveland dataset was partitioned into two disjoint subsets. A subset of 153 patient records was used for Bayesian network structure learning and model training within the cross-validation procedure, while the remaining records were reserved for synthetic data generation, as detailed in Section 3.6. Following data preprocessing, a Bayesian network was learned from a randomly selected subset of 153 patient records from the Cleveland dataset. To improve biological plausibility and reduce overfitting, structural constraints were imposed during the learning process. An initial whitelist was defined to include all possible directed edges between distinct variables, thereby allowing the structure learning algorithm to explore the full space of admissible network configurations. Domain knowledge was subsequently incorporated through the construction of a blacklist, which excluded edges deemed implausible or clinically inconsistent based on established medical reasoning and causal considerations.

In particular, edges violating basic temporal or biological constraints were prohibited. These included dependencies from outcome or symptomatic variables to non-modifiable demographic factors (e.g., from `HeartDisease` or `ChestPain` to `Age` or `Sex`), as well as connections between clinically unrelated physiological measurements (e.g., `FastingBS` to `ChestPain` or `Sex`).

Such exclusions were defined a priori to prevent spurious associations and reverse causal relationships, in line with established principles of causal modelling in epidemiology [18].

These constraints ensured that the resulting Bayesian network remained largely data-driven while being restricted to structures that are clinically interpretable and consistent with domain expertise. The network structure was learned using the Hill-Climbing (HC) algorithm [19], a score-based search method that iteratively applies local modifications to the network topology to maximise a predefined scoring function. In this study, the Bayesian Information Criterion (BIC) was employed as the optimisation criterion, balancing model fit against structural complexity. The resulting network captures conditional dependencies among clinical variables in a manner consistent with both the observed data and prior clinical assumptions.

3.5. Parameter estimation inference and classification

Once the network structure was established, Conditional Probability Tables (CPTs) were estimated using Bayesian parameter learning techniques, following standard formulations for probabilistic graphical models [20,21]. These CPTs quantify the likelihood of each node's states given the configuration of its parent nodes and provide the foundation for probabilistic reasoning and causal interpretation within

Bayesian networks. The resulting model constitutes a complete probabilistic system capable of representing both marginal and conditional dependencies among clinical variables.

The trained Bayesian network was subsequently employed for probabilistic inference on the target variable (`HeartDisease`) given new or partial evidence from patient attributes. This inferential capacity was leveraged for classification tasks by predicting the most probable disease state for unseen instances, aligning with recent applications of explainable Bayesian networks in clinical decision-support systems [22]. Both internal validation (via cross-validation on the Cleveland dataset) and external validation (using the Hungarian and Switzerland datasets) were performed to assess predictive accuracy and model generalisability.

3.6. Synthetic data overview

To further explore the role of generative modelling in clinical data analysis [23,24], synthetic medical records were generated using a Bayesian network trained on real patient data. Each synthetic record represents a single patient-level, cross-sectional clinical profile, encoded in tabular form and described by the same set of 13 clinical variables and the `HeartDisease` outcome used in the original datasets. This strategy enables the simulation of clinically plausible data distributions while increasing the effective training set size without requiring additional real samples.

The Cleveland dataset, comprising 303 patient records, was partitioned into two disjoint subsets with distinct roles in the experimental workflow. A first subset of 153 records was used for Bayesian network structure learning and model training within the cross-validation procedure. The remaining 150 records were exclusively employed to train a separate Bayesian network for synthetic data generation. This design ensures a strict separation between real samples used for predictive modelling and those used for data generation, thereby preventing any potential information leakage and preserving the validity of the experimental evaluation.

Synthetic instances were generated via ancestral sampling from the Bayesian network trained on the excluded real records, using the same structural constraints (whitelist and blacklist) adopted in the primary model to preserve interpretability and clinical coherence. As a result, the synthetic records mirror the structure and variable composition of the original patient data, differing only in their generative origin rather than in their semantic content.

Following generation, synthetic samples were evaluated both independently and in combination with real data for downstream experiments. A performance-based filtering procedure was applied to retain only synthetic instances that preserved or improved predictive performance when integrated with real samples. Specifically, synthetic data were generated as a complete dataset using the Bayesian network trained on the held-out subset of Cleveland records reserved for data generation. The resulting synthetic dataset was then combined with the real training data, and its impact on predictive performance was assessed within the cross-validation framework using a validation set derived from the training data within the cross-validation framework. Predictive performance was measured in terms of F1-score. The synthetic dataset was retained only when its inclusion did not reduce or improve the F1-score compared with the baseline model trained exclusively on real data. This evaluation procedure was performed exclusively within the training process and did not involve the external test cohorts (Hungarian or Switzerland), thereby avoiding information leakage and reducing the risk of selection bias. The filtered synthetic data were subsequently combined with the fixed set of 153 real records to form the augmented training configuration, hereafter referred to as the *Real + Synthetic (filtered)* setting.

The following subsections describe the synthetic sampling procedure and the evaluation of synthetic data quality through principal component analysis (PCA).

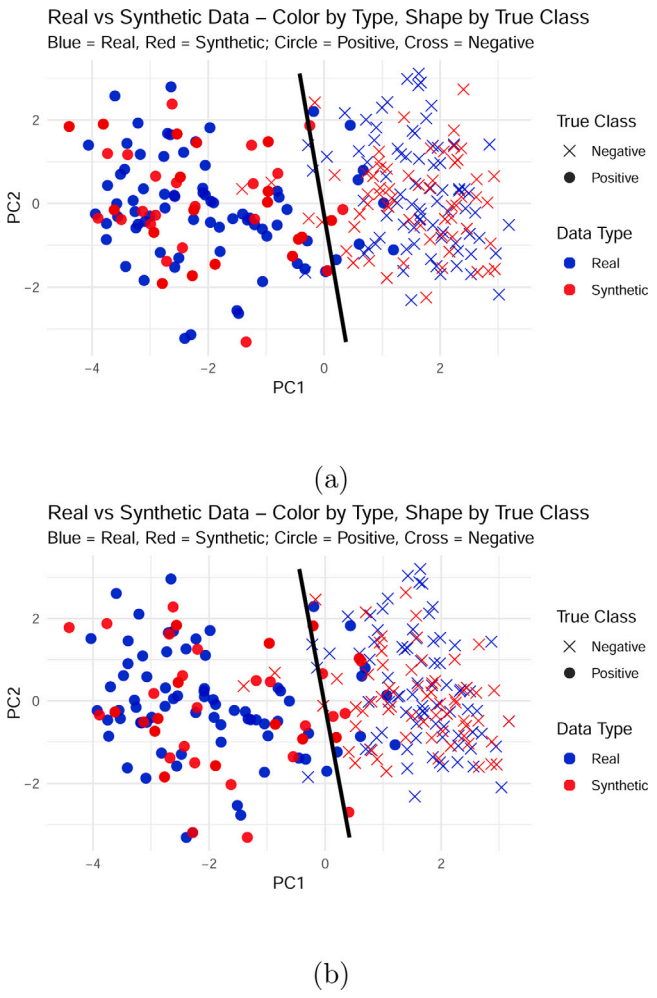

Fig. 3. Projection of real (blue) and synthetic (red) samples onto the first two principal components (PC1 and PC2). Data points are shaped according to the true class (circle = positive, cross = negative). The black line represents the logistic regression decision boundary. The overlap between positive and negative instances reflects both the intrinsic complexity of the data and information loss due to dimensionality reduction, rather than misclassification. The substantial overlap between real and synthetic samples indicates that the generated data follow a similar distribution to the original data.

3.6.1. Synthetic sampling and quality assessment

Synthetic records were generated by sampling from the joint probability distribution encoded by the Bayesian network fitted on excluded real data. This process ensured that the generated instances reflected both the marginal and conditional dependencies observed in the original dataset. To account for generative variability, multiple synthetic datasets were produced and analysed independently across experiments. To qualitatively evaluate the fidelity of the synthetic data, Principal Component Analysis (PCA) was conducted on the combined set of real and synthetic records (see Fig. 3). The projection onto the first two principal components revealed substantial overlap between the two groups, indicating alignment in their underlying structure. A logistic regression classifier was also fitted in the PCA space to approximate the decision boundary, providing additional insight into the separability of synthetic and real samples. The overlap between real and synthetic samples observed in Fig. 3 indicates that the synthetic data follow a similar distribution to the real data, suggesting that they can be meaningfully integrated with real samples for predictive modelling.

3.7. Experimental design

To evaluate the impact of synthetic data augmentation, a comprehensive experimental pipeline was implemented. The overall procedure, summarised in Algorithm 1, combines internal cross-validation and external validation across multiple datasets.

The analysis considers three training configurations: (i) real data only (baseline), (ii) synthetic data only, and (iii) a hybrid dataset combining real and synthetic samples. For each configuration, predictive performance is assessed using 10-fold cross-validation on internal data and subsequently evaluated on two independent external cohorts.

Specifically, a subset of the Cleveland dataset is used for model development, and generalisation performance is evaluated on the Hungarian and Switzerland datasets. This experimental design enables a systematic comparison of robustness and predictive performance under different data augmentation strategies.

Algorithm 1 Experimental pipeline for heart disease prediction

Require: Cleveland dataset C, Hungarian dataset H, Switzerland dataset S
Ensure: Performance metrics for internal and external validation
Preprocess all datasets by imputing missing values and discretising continuous variables
Split C into two disjoint subsets: C_{train} with 153 records and C_{gen} with 150 records
Learn a Bayesian network structure from C_{train} using clinically informed whitelist and blacklist constraints
Fit the Bayesian network parameters on C_{train}
Learn a separate Bayesian network from C_{gen} for synthetic data generation
Generate synthetic dataset C_{synth} using ancestral sampling
Initialize $cv\_metrics\_internal \leftarrow \emptyset$
Initialize $cv\_metrics\_external \leftarrow \emptyset$
Initialize $cv\_metrics\_external\_synth \leftarrow \emptyset$
for $i = 1 : 10$ **do**
  $fold\_train \leftarrow train\_cleveland \setminus folds[i]$
  $fold\_test \leftarrow train\_cleveland[folds[i]]$
  Baseline - Internal validation
  $res\_internal \leftarrow evaluate\_model(bn\_model, fold\_test,$ '$Cleveland\_Foldi$')
  $cv\_metrics\_int \leftarrow cv\_metrics\_int \cup res\_int$
  Synthetic data validation
  $res_{hungarysynth} \leftarrow evaluate\_model(bn\_model, hun\_synth)$
  $res_{switzerland\_synth} \leftarrow evaluate\_model(bn\_model, switz\_synth,)$
  $cv\_metrics\_synth \leftarrow cv\_metrics\_ext\_synth \cup \{res\_hung\_synth, res\_switz\_synth\}$
  Real and synthetic data validation
  $res_{hungary} \leftarrow$ evaluated_model(bn_model, $hungary$,'Hungary_Fold$i$')
  save_pred(bn_model, $hungary$,'Hungary_Fold$i$')
  $res_{switzerland} \leftarrow$ evaluated_model(bn_model, $switzerland$, "Switzerland_Fold$i$")
  save_pred(bn_model, $switzerland$, "Switzerland_Fold$i$")
  $cv\_metrics\_ext \leftarrow cv\_metrics\_ext \cup \{reshungary, res_{switzerland}\}$
end for
return $\{cv\_metrics\_int, cv\_metrics\_ext, cv\_metrics\_ext\_synth\}$
Compare performance across the three configurations using the collected metrics
return Internal and external validation metrics

3.8. Performance evaluation

To assess the validity of synthetic data generation, a comparative analysis was conducted by tracking predictions across models trained with and without synthetic augmentation. Each patient-level outcome was categorised as improved, worsened, or unchanged depending on prediction consistency and correctness relative to the baseline.

Model performance was evaluated using standard classification metrics, including *accuracy*, *precision*, *recall*, and *F1-score*. These metrics were computed for each fold and averaged across internal and external test sets.

Statistical analyses were performed to compare the performance distributions between the baseline and augmented models to ascertain the statistical significance of the improvements induced by synthetic data.

4. Results and discussion

The following section presents the results in line with the study objectives (see Section 3).

Specifically, we report on the impact of synthetic data generation on predictive performance (Research Question 1), the influence of real data quality on the effectiveness of synthetic augmentation (Research Question 2), and the role of random sampling in shaping model outcomes (Research Question 3)

4.1. Predictive performance with synthetic augmentation

The comparative analysis of the three training configurations, real data only (baseline), real combined with synthetic, and synthetic only, revealed systematic differences in external validation performance. Models trained solely on real data established a moderate baseline, with mean F1-scores of approximately 0.54 in Hungary and 0.84 in Switzerland.

When synthetic data were added to the real dataset, performance improved consistently in both cohorts. The mean F1-score increased to 0.57 in the Hungary Dataset and 0.88 in the Switzerland Dataset, accompanied by gains in accuracy (0.51 → 0.58 in the Hungary Dataset; 0.73 → 0.78 in the Switzerland Dataset), indicating that synthetic augmentation contributed to more robust and generalisable predictions.

In contrast, models trained exclusively on synthetic datasets performed above chance levels but did not achieve the improvements observed with the mixed configuration. In the Hungary cohort, the mean F1-score decreased to 0.52, while in the Switzarland cohort, performance remained comparable to the baseline (F1 → 0.81). These findings indicate that synthetic records alone cannot fully replace real data, but may preserve useful predictive information.

For completeness, the results of all evaluation metrics, including accuracy, precision, recall, and F1-score, are reported in Table 3.

4.2. Impact of real data quality on synthetic data effectiveness

To address RQ2, we investigated how the quality of the real data used to generate synthetic samples affects predictive performance. We hypothesised that synthetic data derived from well-structured, complete datasets would yield more meaningful improvements than data generated from noisier or less consistent sources [25].

Fig. 4 summarises the comparison across the Hungary and Switzerland cohorts, using accuracy, precision, and F1-score as evaluation metrics. Each plot contrasts three training configurations: models trained on real data only (Base), on synthetic data only (Synth only), and on a combination of real and filtered synthetic data (Real + Synth filtered).

Across both cohorts, models trained on the combined dataset generally outperformed those based solely on real data. The improvements were particularly marked in accuracy and precision, indicating that filtered synthetic augmentation enhances model generalisation and reduces false positives [26]. For both cohorts, ANOVA [27] results confirmed statistically significant effects (e.g., F(2,27) = 23.06,[1] $\eta^2 = 0.63$[2]

[1] The F-statistic is the test statistic used in analysis of variance (ANOVA) to assess whether there are statistically significant differences between group means.

Table 3
External validation results (mean values across folds) for the three training configurations on the Hungarian (H) and Swiss (S) cohorts.

Configuration	Hungarian (H)				Switzerland (S)			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
Real only	0.506	**0.810**	0.410	0.540	0.732	0.764	0.930	0.840
Real + Synthetic	**0.580**	0.780	**0.450**	**0.570**	**0.780**	**0.820**	**0.940**	**0.880**
Synthetic only	0.560	0.670	0.430	0.520	0.660	0.700	0.930	0.800

Table 4
External validation metrics for synthetic datasets (DS1–DS5) on the Hungarian (H) and Swiss (S) cohorts, compared with the real-only baseline.

Dataset	Hungarian (H)				Switzerland (S)			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
Real only	**0.507**	**0.817**	**0.408**	**0.544**	**0.733**	**0.764**	**0.938**	**0.842**
DS1	0.557	0.767	0.434	0.555	0.778	0.816	0.939	0.873
DS2	0.552	0.781	0.433	0.557	0.786	0.824	0.939	0.878
DS3	0.560	0.728	0.434	0.544	0.719	0.749	0.937	0.856
DS4	0.571	0.760	0.515	0.545	0.797	0.830	0.946	0.884
DS5	0.551	0.775	0.431	0.554	0.789	0.827	0.940	0.880

for precision in Switzerland; $F(2, 27) = 41.56$, $\eta^2 = 0.75$ in Hungary), highlighting the robustness of these differences.

When comparing the two external validations, distinct patterns emerge. In the Switzerland cohort, which exhibits a lower proportion of missing values and more consistent variable availability (see Tables 1 and 2), the inclusion of synthetic data led to consistently high and statistically significant gains across all three metrics. In contrast, the Hungarian cohort, characterised by higher levels of missingness and greater data heterogeneity, exhibited weaker or partially non-significant improvements, particularly for the F1-score. The observed variability in performance across datasets can be attributed to differences in data quality and completeness. In particular, the Switzerland dataset, despite containing some missing values, exhibits more consistent structure and variable availability, allowing the synthetic data to better capture meaningful patterns. In contrast, the Hungarian dataset is characterised by a higher level of missingness and greater heterogeneity, which may limit the ability of the generative model to learn reliable dependencies and reduce the effectiveness of synthetic augmentation. Synthetic samples derived from well-curated datasets, such as Switzerland, are associated with measurable performance gains, whereas those generated from noisier datasets offer limited benefit. This highlights the crucial role of data quality in determining the success of synthetic data approaches in biomedical predictive modelling [28].

4.3. Impact of random subsampling on synthetic data variability

To assess the impact of random subsampling on the variability of synthetic data generation (RQ3), we considered random sampling as an additional technique. The objective was to assess how the stochasticity of the sampling process propagates to the generation of synthetic records and, ultimately, affects predictive performance. Specifically, five Synthetic Datasets (DS1–DS5) were generated by repeatedly sampling distinct random subsets of 150 instances from the Cleveland dataset. Each subset was used to train an independent Bayesian network under identical structural constraints, which was subsequently employed to generate 150 synthetic records via ancestral sampling.

Although all synthetic datasets were generated from the same original data source and followed the same modelling procedure, differences in the sampled real subsets introduce variability in the learned network parameters and, consequently, in the resulting synthetic data distributions. Each synthetic dataset was therefore evaluated independently through external validation on the Hungarian and Switzerland cohorts, allowing the assessment of performance variability induced solely by random subsampling.

To assess how the quality of the real data directly affects the quality of the synthetic data generated from it, we considered random sampling as an additional technique. The objective was to assess how the stochasticity of the sampling process propagates to the generation of synthetic records and, ultimately, affects predictive performance. Specifically, to assess the effect of real data variability on the quality of synthetic data and their subsequent predictive performance, we generated five synthetic datasets, each trained on a randomly sampled subset of 150 instances from the Cleveland dataset. Each subset was used to train a distinct Bayesian network, which was then employed to generate 150 synthetic records. The synthetic data were evaluated through 10-fold cross-validation, using the external validation.

The results indicate that randomisation introduces a non-negligible source of variability. In several runs, the synthetic datasets maintained well-defined class boundaries, resulting in predictive performance comparable to that of models trained on real data. In other cases, however, random sampling produced noisier synthetic distributions, resulting in higher misclassification rates and decreased predictive accuracy. These findings indicate that random subsampling constitutes a non-negligible source of variability in synthetic data generation and downstream predictive performance (see Table 4). The PCA projections provide a qualitative illustration of the variability induced by random subsampling. As shown in Fig. 5, some synthetic datasets were generated from specific real-data subsamples (e.g., DS3) that exhibit weaker class separation and increased overlap in the projected space. In these instances, the absence of distinct decision boundaries directly corresponded to increased prediction inaccuracy and reduced predictive performance.

Conversely, when the sampling process yielded more representative subsets of the real data, the resulting synthetic distributions demonstrated clearer class separation and decision boundaries that more closely resembled those of the original dataset. As shown in Fig. 6, this structured distribution facilitated more stable model training and supported improved generalisation. Further evidence of the predictive contribution of synthetic records is illustrated in Fig. 7, where

[2] Eta squared (η^2) is a measure of effect size in ANOVA, representing the proportion of total variance in the dependent variable explained by the grouping factor.

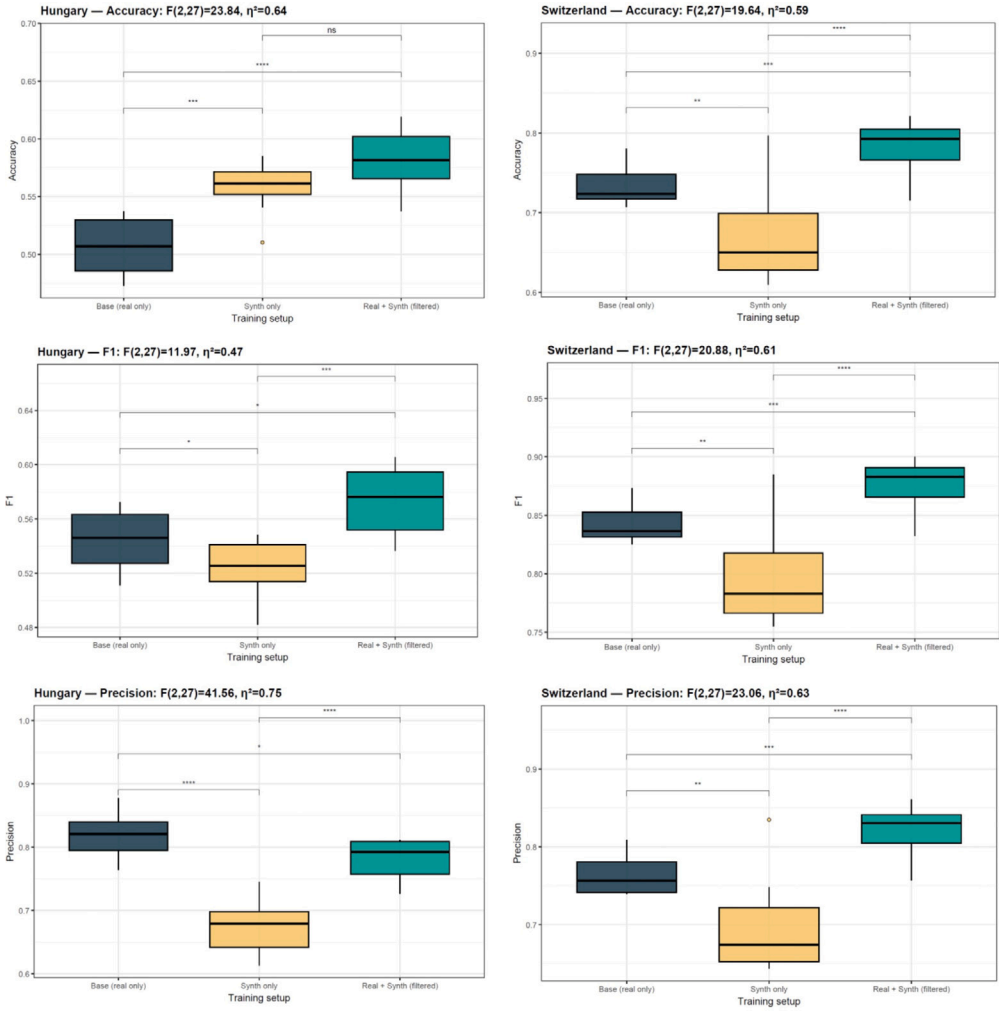

Fig. 4. Impact of real data quality on the effectiveness of synthetic data augmentation. Boxplots show model performance across three training configurations (Base: real only; Synth only; Real + Synth (filtered)) for the Hungary (left column) and Switzerland (right column) cohorts. Panels (a) and (d) report accuracy, (b) and (e) F1-score, and (c) and (f) precision. Each boxplot summarises performance across cross-validation folds. The Real + Synth (filtered) configuration generally shows higher median performance, with more consistent gains observed for the Switzerland cohort. Statistical significance was assessed using one-way ANOVA with Bonferroni-adjusted post hoc comparisons (*p* < 0.05, **p** < 0.01, ***p* < 0.001).

individual instances are categorised based on their effect on model performance. Notably, only a subset of synthetic samples positively influenced the predictions, underscoring the importance of data quality and structure.

5. Conclusions

The findings of this study demonstrate that synthetic data generated through Bayesian networks can enhance predictive modelling for heart disease when appropriately integrated with real clinical records. The inclusion of synthetic samples improved model accuracy and precision across the experimental configurations, with particularly consistent gains observed for the Switzerland cohort, where data completeness and structure supported more reliable augmentation. These results suggest that generative Bayesian modelling can effectively help mitigate data scarcity while maintaining interpretability and clinical plausibility,although its effectiveness depends on the characteristics of the underlying data.

Building on these results, the study provided a rigorous evaluation of how the quality of real data determines the effectiveness of synthetic augmentation. By combining probabilistic graphical modelling with clinically informed structural constraints, we developed a transparent and reproducible framework for generating, validating, and assessing synthetic data externally. The comparative analysis across the Cleveland, Hungary, and Switzerland datasets revealed that performance improvements are strongly influenced by modulated by the integrity and coherence of the underlying data, reinforcing that the reliability of generative models ultimately depends on the quality of their training sources.

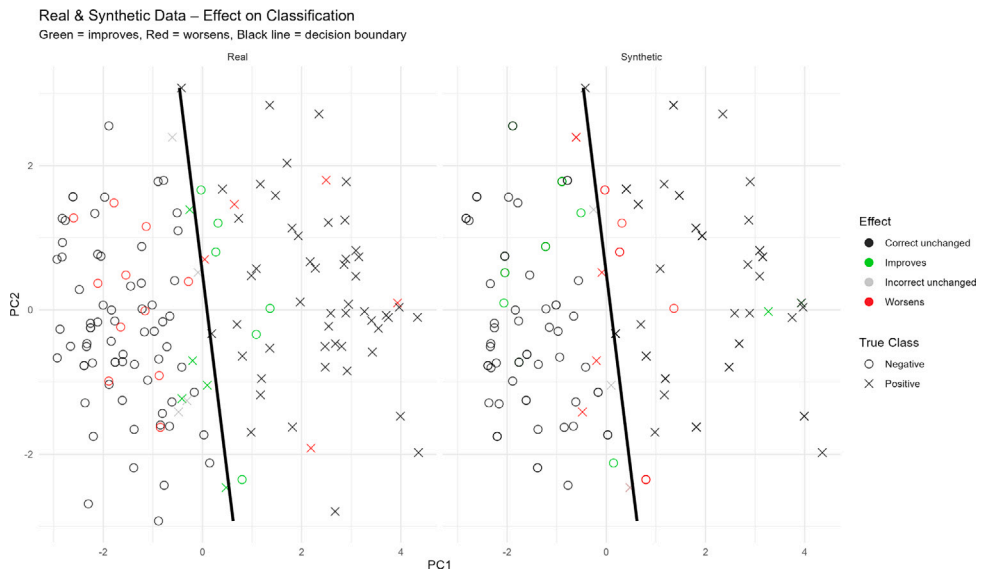

Fig. 5. PCA projection of a synthetic dataset (DS3) generated from a poorly representative real data subset. Data points are projected onto the first two principal components (PC1 and PC2). Symbols denote the true class (circle = negative, cross = positive). Colours indicate the effect of each instance on predictive performance relative to the real-only baseline (green = improves, red = worsens, grey = incorrect unchanged, black = correct unchanged). The black line represents the logistic regression decision boundary fitted in the PCA space.

Fig. 6. PCA projection of synthetic dataset DS4 generated from a representative subset. Data points are projected onto the first two principal components (PC1 and PC2). Symbols denote the true class (circle = negative, cross = positive). Colours indicate the effect of each instance on predictive performance relative to the real-only baseline (green = improves, red = worsens, grey = incorrect unchanged, black = correct unchanged). The black line represents the logistic regression decision boundary fitted in the PCA space.

Importantly, the effectiveness of the proposed framework is not fixed, but depends on several methodological design choices. In particular, model performance may vary according to the size and representativeness of the real subset used for generative training, the structural constraints imposed during network learning (whitelist and blacklist specification), the number of synthetic samples generated, and the filtering criteria applied before data augmentation. The variability observed under random subsampling further demonstrates that different real-data partitions can lead to different learned parameters and, consequently, to synthetic datasets with distinct distributional properties and predictive impact. These findings indicate that synthetic augmentation should be regarded as a controlled modelling strategy whose benefits depend on careful calibration of its generative components.

A limitation of this work concerns the relatively small sample size of the UCI Heart Disease datasets, which contain between 100 and 300 patient records. Such limited cohorts inevitably restrict statistical power and generalisability. Future research should systematically investigate the framework's sensitivity to key generative parameters and structural assumptions, and evaluate its robustness on larger, more heterogeneous clinical datasets. In doing so, Bayesian network–based synthetic data generation could become a reliable and transparent tool to mitigate data scarcity in clinical research and to support trustworthy artificial intelligence in healthcare.

CRediT authorship contribution statement

Ilaria Lazzaro: Writing – original draft, Validation, Methodology, Investigation. **Marianna Milano:** Writing – review & editing. **Allan**

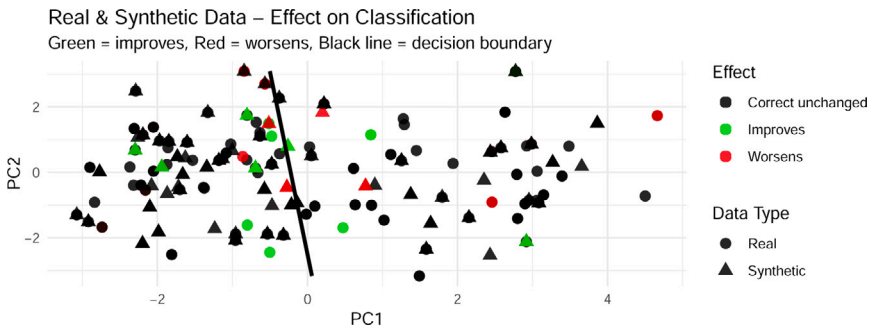

Fig. 7. PCA projection illustrating the effect of individual synthetic instances on model predictions. Triangles represent synthetic samples and circles represent real samples. Colours indicate their impact relative to the real-only baseline (green = improves, red = worsens, black = correct unchanged). The black line denotes the logistic regression decision boundary.

Tucker: Writing – review & editing, Validation, Supervision. **Mario Cannataro:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been partially supported by the OFIDIAPlus (Operational Fire Danger preventIon plAtform Plus) project under the INTERREG GREECE-ITALY 2021–2027 PROGRAMME.

Data availability

All datasets used in this work are freely available in the UCI repository.

References

[1] A.L. Beam, I.S. Kohane, Big data and machine learning in health care, JAMA 319 (13) (2018) 1317–1318.
[2] J. Ye, et al., The role of artificial intelligence for the application of integrated electronic health records, NPJ Digit. Med. 7 (1) (2024) 45.
[3] M. Vaduganathan, et al., The global burden of cardiovascular diseases and risk, J. Am. Coll. Cardiol. 80 (25) (2022) 2347–2360.
[4] M. Gong, X. Sun, H. Li, K Zhang, Synthetic data in machine learning for medicine and healthcare, Nat. Biomed. Eng. 5 (5) (2021) 493–497.
[5] M. Lappenschaar, A. Hommersom, P.J.F. Lucas, J. Lagro, S Visscher, Multilevel Bayesian networks for the analysis of hierarchical health care data, Artif. Intell. Med. 57 (3) (2013) 171–183.
[6] V.C. Pezoulas, D.I. Zaridis, E. Mylona, C. Androutsos, K. Apostolidis, N.S. Tachos, D.I. Fotiadis, Synthetic data generation methods in healthcare: A review on open-source tools and methods, Comput. Struct. Biotechnol. J. 23 (2024) 2892–2910, http://dx.doi.org/10.1016/j.csbj.2024.07.005.
[7] D. Kaur, J.C. Issac, A Patel, Application of Bayesian networks to generate synthetic health data, JAMIA Open 3 (4) (2020) 628–633.
[8] J. Young, P. Graham, R. Penny, Using Bayesian networks to create synthetic data, J. Off. Stat. 25 (4) (2009) 549–567.
[9] M. Kelly, R. Longjohn, K Nottingham, The UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences, 2014, Available at: https://archive.ics.uci.edu.
[10] R. Detrano, Cleveland heart disease dataset, UCI Mach. Learn. Repos. (1989) Available at: https://archive.ics.uci.edu/ml/datasets/Heart+Disease.
[11] A. Janosì, Hungarian heart disease dataset, UCI Mach. Learn. Repos. (1989) Available at: https://archive.ics.uci.edu/ml/datasets/Heart+Disease.
[12] M. Pfisterer, Switzerland heart disease dataset, UCI Mach. Learn. Repos. (1989) Available at: https://archive.ics.uci.edu/ml/datasets/Heart+Disease.
[13] P.K. Whelton, et al., 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC /NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults, Hypertension 71 (2018) e13–e115, http://dx.doi.org/10.1161/HYP.0000000000000065.
[14] Centers for Disease Control and Prevention (CDC), About high cholesterol, 2025, https://www.cdc.gov/cholesterol/about/index.htm, (Accessed 2025).
[15] D.M. Maslove, T. Podchiyska, H.J. Lowe, Discretization of continuous features in clinical datasets, J. Am. Med. Informatics Assoc. 20 (3) (2013) 544–553, http://dx.doi.org/10.1136/amiajnl-2012-000929.
[16] J. Dougherty, R. Kohavi, M. Sahami, Supervised and unsupervised discretization of continuous features, in: Proceedings of the 12th International Conference on Machine Learning, ICML-95, Morgan Kaufmann Publishers, Inc., San Francisco, CA, USA, 1995, pp. 194–202, San Francisco, CA, USA, 1995.
[17] F. Nojavan, H. Smith, et al., Comparative analysis of discretization methods in Bayesian network modelling, Env. Model. Softw. 95 (2017) 1–12.
[18] S. Greenland, J. Pearl, J.M. Robins, Causal diagrams for epidemiologic research, Epidemiology 10 (1) (1999) 37–48.
[19] J. Więckowski, B. Kizielewicz, J. Kołodziejczyk, Application of hill climbing algorithm in determining the characteristic objects preferences based on the reference set of alternatives, in: I. Czarnowski, R. Howlett, L. Jain (Eds.), Intelligent Decision Technologies. IDT 2020. Smart Innovation, Systems and Technologies, 193, Springer, Singapore, 2020, http://dx.doi.org/10.1007/978-981-15-5925-9_29.
[20] M. Scutari, J.-B. Denis, Bayesian Networks: With Examples in R, Chapman & Hall/CRC Press, Boca Raton, FL, USA, 2021.
[21] J. Kuipers, P. Suter, G. Moffa, Efficient sampling and structure learning of Bayesian networks, J. Comput. Graph. Statist. 31 (3) (2022) 639–650.
[22] J.M. Ordovas, D. Rios Insua, A. Santos-Lozano, A. Lucia, A. Torres, A. Kosgodagan, J.M. Camacho, A Bayesian network model for predicting cardiovascular risk, Comput. Methods Programs Biomed. 231 (2023) 107405, http://dx.doi.org/10.1016/j.cmpb.2023.107405.
[23] Y. Liang, B. Nobakht, G. Lindsay, The application of synthetic data generation and data-driven modelling in the development of a fraud detection system for fuel bunkering, Meas.: Sensors 18 (2021) 100225, http://dx.doi.org/10.1016/j.measen.2021.100225.
[24] V.C. Pezoulas, D.I. Zaridis, E. Mylona, C. Androutsos, K. Apostolidis, N.S. Tachos, D.I. Fotiadis, Synthetic data generation methods in healthcare: A review on open-source tools and methods, Comput. Struct. Biotechnol. J. 23 (2024) 2892–2910, http://dx.doi.org/10.1016/j.csbj.2024.07.005.
[25] Z. Qian, T. Callender, B. Cebere, S.M. Janes, N. Navani, M. van der Schaar, Synthetic data for privacy-preserving clinical risk prediction, Sci. Rep. 14 (1) (2024) 15455, Available online: http://dx.doi.org/10.1038/s41598-024-72894-y. (Accessed 15 October 2025).
[26] J. Wang, Y. Li, X. Chen, H. Zhao, Y. Zhang, Synthetic data for healthcare AI: Balancing fidelity and privacy, NPJ Digit. Med. 7 (3) (2024) 112, Available online: http://dx.doi.org/10.1038/s41746-024-00945-2. (Accessed 15 October 2025).
[27] A. Benavoli, G. Corani, J. Demšar, M. Zaffalon, Time for a change: A tutorial for comparing multiple classifiers through Bayesian analysis, J. Mach. Learn. Res. 18 (77) (2017) 1–36.
[28] R.J. Chen, M.Y. Lu, T.Y. Chen, D.F. Williamson, F. Mahmood, Synthetic data in machine learning for medicine and healthcare, Nat. Biomed. Eng. 5 (6) (2021) 493–497.

Further reading

[1] A. Janosi, W. Steinbrunn, M. Pfisterer, R. Detrano, Heart disease, UCI Mach. Learn. Repos. (1989) http://dx.doi.org/10.24432/C52P4X.
[2] A.M. Alaa, B. van Breugel, E. Saveliev, M. van der Schaar, How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models, 2021, CoRR vol. . Available: https://arxiv.org/abs/2102.08921.