information

MDPI

Article

Synthetic Data Generation for Binary and Multi-Class Classification in the Health Domain

Camila Guerreiro [1], Fátima Leal [1,*] and Micaela Pinho [1,2,3]

[1] Research on Economics, Management and Information Technologies, REMIT, Portucalense University, 4200-072 Porto, Portugal; 50716@alunos.upt.pt (C.G.); michaelapinho@hotmail.com (M.P.)
[2] Instituto Jurídico Portucalense, IJP, Portucalense University, 4200-072 Porto, Portugal
[3] Research Unit in Governance, Competitiveness and Public Policy, GOVCOPP, Aveiro University, 3810-193 Aveiro, Portugal
* Correspondence: fatimal@upt.pt

Abstract

The growing demand for data-driven solutions in healthcare is often hindered by limited access to high-quality datasets due to privacy concerns, data imbalance, and regulatory constraints. Synthetic data generation has emerged as a promising strategy to address these challenges by creating artificial yet statistically valid datasets that preserve the underlying patterns of real data without compromising patient confidentiality. This study explores methodologies for generating synthetic data tailored to binary and multi-class classification problems within the health domain. We employ advanced techniques such as probabilistic modelling, generative adversarial networks, and data augmentation strategies to replicate realistic feature distributions and class relationships. A comprehensive evaluation is conducted using benchmark healthcare datasets, measuring fidelity, diversity, and utility of the synthetic data in downstream predictive modelling tasks. The original dataset consisted of 2125 imbalanced cases, both in the binary and multi-class classification scenarios. Experimental results demonstrate that models trained on synthetic datasets achieve performance levels comparable to those trained on real data, particularly in scenarios with severe class imbalance. The findings underscore the potential of synthetic data as a privacy-preserving enabler for robust machine learning applications in healthcare, facilitating innovation while adhering to strict data protection regulations.

Keywords: synthetic data; binary; multi-class; classification; health; data balancing

Academic Editors: Francesco Isgrò, Huiyu Zhou and Daniele Ravi

Received: 16 September 2025
Revised: 4 November 2025
Accepted: 12 November 2025
Published: 14 November 2025

Citation: Guerreiro, C.; Leal, F.; Pinho, M. Synthetic Data Generation for Binary and Multi-Class Classification in the Health Domain. *Information* **2025**, *16*, 986. https://doi.org/10.3390/info16110986

1. Introduction

The use of Machine Learning (ML) in healthcare is increasingly dependent on the availability of high-quality data. However, access to sufficiently large and representative datasets remains a persistent challenge due to privacy constraints, ethical considerations, and the high cost of medical data collection. Moreover, healthcare data often suffers from class imbalance, where rare but clinically critical outcomes are underrepresented, making it difficult for ML models to learn robust and generalizable decision rules. These challenges highlight the need for alternative approaches that enable the development of reliable predictive models while preserving patient confidentiality and ensuring statistical validity.

Synthetic data generation has emerged as a powerful solution to these limitations. By creating artificial data that replicates the statistical properties of real-world datasets, synthetic data offers a way to overcome scarcity, mitigate class imbalance, and reduce the risk of overfitting. In contrast to simple data augmentation, synthetic data generation

maintains both the marginal distributions of variables and the complex interdependencies between them, which are crucial for healthcare applications. Preserving these relationships ensures that ML models can capture meaningful patterns without compromising the integrity of predictions. Recent advances in generative modelling, including probabilistic frameworks, Bayesian networks, and neural-based approaches, have further strengthened the ability to produce synthetic data that is both realistic and diverse [1,2].

In the specific context of patient triage, these challenges are particularly pronounced. Triage datasets are often small, as they rely on structured questionnaires or medical assessments collected under resource-constrained conditions. For this study, the initial dataset comprised 2125 responses from a triage questionnaire aimed at assessing patient care priority. While valuable, this dataset was insufficient to train models capable of generalization to diverse real-world scenarios. To address this, we employed synthetic data generation as a central step in our methodology, ensuring the expansion of the dataset while maintaining its statistical integrity.

The proposed approach is guided by three key objectives. First, preservation of complex relationships between variables such as age, marital status, education level, perceived health, and beliefs about triage criteria, as these interdependencies critically affect predictive accuracy. Second, mitigation of overfitting, by expanding the dataset and providing greater variability in training samples, thereby enhancing generalization capacity. Third, support for model robustness, enabling ML models to handle diverse patient profiles and improve decision reliability in clinical practice. Together, these objectives ensure that the synthetic dataset not only mimics real-world variability but also serves as a reliable foundation for predictive modelling in sensitive healthcare applications.

To achieve these goals, we implemented the Synthetic Data Vault (SDV) framework, which combines advanced probabilistic modelling and ML to generate synthetic data that faithfully reflects the statistical structure of the original dataset [3–5]. The SDV approach was selected for its ability to automatically detect variable types, capture complex correlations, and scale efficiently to large volumes of synthetic data. By leveraging Gaussian Copula models within SDV, we ensured that the generated data preserved both distributions and dependencies, enabling the training of robust models for binary and multi-class triage classification.

Finally, to address the problem of class imbalance, we explicitly integrated data balancing into the generation process. Two classification schemes were considered: a binary model distinguishing between low- and high-priority patients, and a multi-class model offering finer stratification across four priority levels. In both cases, synthetic data was generated to balance class distributions, reducing bias and enhancing predictive performance. The statistical similarity between real and synthetic data was evaluated using a comprehensive set of descriptive measures (e.g., mean, standard deviation, skewness, kurtosis), complemented by visual inspection of distributions. Iterative refinement ensured that the synthetic dataset aligned closely with the original data, thereby providing a statistically consistent yet diversified foundation for predictive modelling.

This work proposes and evaluates a synthetic data generation methodology tailored to healthcare triage applications. By combining SDV-based generative models with explicit class balancing, we demonstrate how synthetic data can enhance the development of robust ML models in both binary and multi-class classification tasks. The contributions of this study are threefold: (*i*) systematic methodology for generating and validating synthetic healthcare data that preserves statistical properties and inter-variable relationships; (*ii*) the application of SDV and Gaussian Copula modelling to expand small triage datasets while addressing class imbalance; and (*iii*) a demonstration of the effectiveness of synthetic data in supporting reliable ML-based decision making for patient prioritization.

2. Related Work

In healthcare, datasets are often imbalanced and eliciting societal data concerning their support for the criteria that should guide patient prioritization are often imbalanced and biassed. Therefore, the synthetic data generation offers a promising alternative: it preserves the statistical patterns of real data without exposing individual records [6]. In this context, synthetic datasets accurately mimic joint distributions and structural dependencies, enabling simulation, behavioural modelling and algorithm development when real samples are scarce or legally restricted [7,8]. Synthetic data are used to address (*i*) data scarcity; and (*ii*) privacy and regulation.

Modern ML and deep learning algorithms require large amounts of data to achieve robust performance. In medicine, rarity of certain conditions and logistical or ethical hurdles in data collection limit sample size. Synthetic generation enlarges training corpora while retaining key statistical properties, boosting predictive accuracy even in small-sample scenarios [8]. In turn, health data are highly sensitive and regulated [9]. Even with anonymisation, re-identification risks remain. High fidelity synthetic data, decoupled from real patients, facilitate research and innovation within ethical and legal boundaries [6,8].

Access to high quality and diverse healthcare data is often limited due to privacy concerns, regulatory constraints, and the rarity of certain clinical scenarios. To address these challenges, synthetic data generation has emerged as a valuable technique for augmenting datasets, enabling model training, testing, and validation without compromising patient confidentiality. Early work relied on classical statistical techniques; recent years have seen a shift toward ML based generative models capable of learning complex variable dependencies and producing novel combinations ("sampled zeros") [6,7,10]. These techniques aim to produce realistic and representative data that preserve the statistical properties of the original dataset, thereby supporting robust and generalizable ML models for patient prioritization and other healthcare tasks.

Bayesian-Network (BN) encode conditional relationships through directed acyclic graphs. Sampling from the learned joint distribution allows the generation of synthetic records. Despite their interpretability, they become computationally infeasible as dimensionality increases, and they struggle to model non-linear dependencies, compromising both data quality and privacy [10,11]. Limitations include (*i*) exponential growth in learning complexity with the number of variables [3]; (*ii*) poor representation of highly non-linear relationships [11]; and (*iii*) tendency to memorize real data, putting privacy at risk [11].

Neural Generative Models such as Variational Autoencoders (VAE) and Generative Adversarial Networks (GAN) have gained prominence. VAE map data to a latent space through variational inference, being effective with continuous variables, including longitudinal and eye-tracking datasets [6,7]. However, they struggle with categorical attributes and require significant computational resources for large datasets [4,11]. GAN, on the other hand, combine a generator and a discriminator in an adversarial training dynamic. For tabular data, variants such as Conditional Tabular Generative Adversarial (CTGAN) apply conditional techniques to handle categorical variables and class imbalances, improving synthetic fidelity and predictive performance when combined with real data [4,8].

Despite their potential, neural models face significant challenges in sensitive domains such as healthcare. They often operate as black-boxes, making it difficult to trace and justify the generation process, which harms transparency and regulatory compliance [11]. Additionally, problems such as "mode collapse" where the generator produces low diversity data can result in synthetic datasets that omit rare but clinically significant patterns [11]. These models impose a considerable computational burden, require careful hyperparameter tuning, and are sensitive to training conditions, leading to long training times and increased risk of runtime errors when working with large datasets [3,4,11].

Synthia is an open source multidimensional synthetic data generator implemented in Python with version 3.13 [12]. It uses copula-based models, which allow the statistical properties of the observed data to be captured in terms of both individual behaviour and interdependencies between variables [12]. Synthia is further supported by Functional Principal Component Analysis (FPCA), an extension of principal component analysis where the data consist of functions rather than vectors [12]. Synthia is a powerful tool for generating multidimensional synthetic data while preserving complex relationships. However, its applicability is limited to offline processing and it may have shortcomings in accurately replicating variable types and distribution functions from the original data [13].

SDV emerges as a robust and modular solution based on copula theory, specifically designed for tabular data. Gaussian Copula is a function that couples multivariate distribution functions to their univariate margins by describing the dependency structure through a multivariate normal distribution applied to transformed uniform variables [10,14]. Univariate modelling is performed using Gaussian Mixture Models (GMM), which represent a probabilistic model composed of multiple Gaussian components, that links univariate marginals through a multivariate normal distribution after transforming the data using Empirical Cumulative Distribution Functions (ECDF) [10,14]. Sklar's Theorem ensures that any multivariate distribution can be decomposed into its marginal distributions and a copula, enabling the reuse of marginals across domains [6,14].

SDV framework offers several advantages that make it particularly suitable for generating healthcare data. First, it demonstrates high statistical fidelity and clinical realism by preserving the structural and statistical properties of the original dataset, including relationships between tables and columns [3,4]. Additionally, SDV ensures the preservation of variability and the inclusion of rare cases and also eliminates the need to choose a specific copula family, providing flexibility to capture complex dependencies [15].

Given the limitations pointed out in traditional synthetic data generation methods, the SDV emerges as a robust and highly adaptable solution. From a privacy perspective, SDV offers enhanced protection since its probabilistic approach prevents memorization of real data points, promoting an implicit form of differential privacy [4,10]. Furthermore, SDV is recognized for its computational efficiency and scalability, processing large datasets with reduced runtime and greater stability compared to models like GAN and VAE [3,4]. Lastly, among the models included in the SDV library, CopulaGAN and GaussianCopulaSynthesizer are particularly noteworthy for their capacity to model complex dependencies in tabular data. CopulaGAN combines the adversarial training dynamics of GANs with copula-based statistical modelling [5,10,15]. Copulas are multivariate functions that allow for the separation of marginal distributions from the dependency structure, enabling the model to better capture non-linear relationships between variables, an essential feature when working with heterogeneous healthcare data [16]. On the other hand, the GaussianCopulaSynthesizer simplifies this approach by assuming a Gaussian copula structure. It transforms variables into a standard normal space using probability integral transforms and models their joint distribution using a multivariate Gaussian copula. This model has demonstrated high effectiveness in reproducing realistic synthetic datasets while maintaining the integrity of variable correlations and supporting both continuous and categorical data [3,5,15].

Although GAN, VAE, Synthia, and BN have merits in specific domains, SDV represents a more robust, explainable, and efficient alternative for synthetic data generation in sensitive contexts such as healthcare [6]. Within the SDV framework, the GaussianCopulaSynthesizer was selected over CopulaGAN due to its stronger statistical grounding, greater transparency, and lower computational complexity [5]. While CopulaGAN leverages adversarial training to capture complex non-linear dependencies, it also introduces

challenges such as increased training instability and reduced interpretability. In contrast, the Gaussian Copula Synthesizer, by assuming a Gaussian copula structure, provides a simpler yet effective model that preserves variable correlations and supports mixed data types with high fidelity [3–5]. This balance of rigour, explainability, and scalability makes SDV and specifically the Gaussian Copula Synthesizer the ideal choice for this dissertation, enabling the generation of realistic, privacy-preserving synthetic data suitable for healthcare predictive modelling [3–5].

3. Proposed Method

Synthetic data generation involves creating synthetic data that replicates the statistical characteristics of the real dataset, addressing issues such as class imbalance or data scarcity [1,2].

The objective of synthetic data generation is to expand the dataset in a controlled and representative manner, ensuring that the models trained on it can generalize effectively to real world scenarios. This process is particularly valuable in cases where obtaining additional real data is costly, time consuming, or impractical [17,18]. By generating high quality synthetic data, we can not only improve the performance and stability of ML models, providing a richer and more balanced foundation for predictive modelling and decision making but also mitigate potential biases and enhance the reliability of subsequent analyses [2]. To ensure the model can learn not only from the distribution of the data but also from the intricate interactions between different factors, it is essential to maintain the correlations between these variables in the generated synthetic data [19]. Data analysis in health presents unique challenges, particularly when the size of available datasets is limited. In this work, we began with an initial dataset of 2125 responses to a structured questionnaire designed to determine the priority of patient care in medical triage contexts. However, to ensure the development of robust predictive models and reduce the risk of overfitting, it was necessary to expand this dataset through synthetic instances. This process enables better generalization and greater reliability in the predictions made by ML models. Several key factors justify the generation of synthetic data:

- Preservation of Complex Relationships. The variables in the dataset, e.g., age, marital status, educational level, perceived health, and beliefs about triage criteria, display critical interrelationships. An effective predictive model must capture these relationships to make accurate and reliable predictions about patient care priority. Synthetic data generation ensures that these statistical dependencies are preserved and adequately represented in the expanded dataset [1].
- Mitigation of Overfitting. Models trained on small datasets often display overfitted patterns, compromising their ability to generalize to new cases. Expanding the dataset introduces greater diversity, promoting statistical robustness and reducing the probability of overfitting [1,2].
- Support for Model Robustness. A larger synthetic dataset allows models to handle greater variability in response patterns, enabling better adaptation to real-world scenarios. This is particularly important in a context where critical decisions, such as the order of patient care in medical triage, directly depend on the reliable predictive capacity of the models [1,18].

These three characteristics preserving relationships, mitigating overfitting, and enhancing robustness were fundamental considerations in both the design of the synthetic data and the selection of data generation methods. By ensuring that the synthetic data accurately represents real world variability while maintaining statistical consistency, we improved the overall effectiveness and reliability of the machine learning models used for patient triage [1,17,19]. The number of synthetic samples (100,000) was chosen after empirical test-

ing with multiple generation sizes (10,000, 50,000, and 100,000). Statistical indicators such as mean, variance, and pairwise correlations stabilized at this level, confirming that further increases would not meaningfully enhance data fidelity. The chosen size thus ensures both statistical representativeness and computational efficiency for subsequent analyses.

To provide a clear overview of the methodological workflow, Figure 1 illustrates the main stages of the proposed synthetic data generation process. The pipeline begins with data preprocessing and feature encoding, followed by a data balancing stage to address class imbalance in both binary and multi-class triage cases. The balanced dataset is then used to train the Gaussian Copula Synthesizer within the SDV framework, which models the joint probability distributions and dependencies among variables. Once trained, the model generates synthetic records that preserve the statistical characteristics of the original dataset. Finally, the generated data are evaluated using descriptive and distributional metrics to ensure fidelity and representativeness before being applied to healthcare prioritization scenarios.

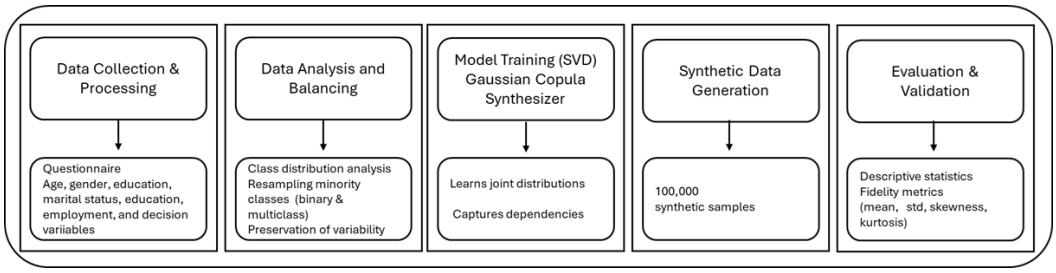

Figure 1. Proposed method.

3.1. SDV

For synthetic data generation, we explored (*i*) Neural Generative Models; (*ii*) Synthia; (*iii*) SDV; (*iv*) BN. Each method was evaluated based on its ability to preserve the statistical properties of the original dataset, handle mixed data types, and generate high quality synthetic data suitable for ML models [6]. The ability to accurately replicate the relationships between the variables, was a key factor in selecting the synthetic data generation method for this project. SDV was ultimately selected due to its superior ability to capture and preserve complex statistical relationships and dependencies between variables [3,4]. Using advanced probabilistic modelling and ML techniques, SDV generates synthetic data that closely mirrors the original dataset's properties [3–5].

The proposed data generation approach offers several advantages crucial for healthcare triage modelling. It automatically detects variable types, whether categorical, continuous, or mixed and learns the relationships between them without manual intervention. This enables the preservation of intricate dependencies, such as correlations between variables [5]. Additionally, the method efficiently scales to generate large volumes of synthetic data while maintaining statistical diversity, ensuring representativeness across patient profiles [1]. Its ability to handle the high dimensionality and complexity of real world healthcare datasets makes it particularly suitable as a reliable foundation for training machine learning models in prioritization contexts [1,2].

By leveraging SDV, ensures that the synthetic data remains statistically representative of the original dataset while providing a scalable and adaptable solution for data augmentation in ML applications [1,2]. The following steps outline the process to generate synthetic data using the SDV framework:

1. Reading and Preprocessing the Data: The real data was initially read from a Comma-separated values (CSV) file, ensuring that all necessary preprocessing steps were applied (e.g., handling missing values, encoding categorical variables where applicable).

2. Metadata Creation: The single table meta data object from the SDV framework was employed to automatically detect the data types and relationships between columns. This was essential to preserve the original characteristics of the data during the synthesis process.
3. Training the Synthesizer: Gaussian Copula model was chosen due to its robustness in modelling complex dependencies between variables [15]. This model was trained on the real dataset to learn the distributions and relationships between the features [2].
4. Generating Synthetic Data: Using the trained Gaussian Copula model, synthetic data were generated. This ensured that the synthetic dataset closely mirrored the statistical properties of the real data, while maintaining the underlying relationships [3,5].
5. Exporting the Synthetic Data: The generated synthetic data was then saved in a CSV file for further analysis, enabling comparisons with the real data to assess the accuracy of the replication process.

3.2. Data Balancing

In ML, data balancing plays a crucial role in ensuring that models can generalize effectively across all classes in a dataset [20]. Class imbalance, where one class significantly outweighs the others in representation, is a common issue, particularly in healthcare datasets where specific outcomes may be rare [21,22]. Imbalanced datasets can lead to models that are biassed toward the majority classes, resulting in poor performance when predicting minority class instances, a critical limitation when dealing with sensitive decisions, such as patient triage prioritization [18,20].

To address this issue, the data generation process explicitly considered class balancing. While real world data often reflects natural imbalances, incorporating balance into the synthetic data generation process was vital for creating a robust training set. This ensured the ML models developed could make accurate predictions for both classes, avoiding bias toward the majority class and enabling better generalization [18,20,23].

This work was developed based on two complementary classification approaches: a binary model, which distinguishes between low priority (class 0) and high priority cases (class 1), and a multi-class model, which provides a more granular stratification across four levels: Not Prioritary (class 0), Low Priority (class 1), Priority (class 2), and High Priority (class 3). These dual perspectives enable both coarse and fine grained prioritization, depending on clinical context and resource availability.

The proposed methodology for balancing the dataset included two key steps:

- Real Data was divided into subsets based on the decision variables. In the binary approach, the aggregated decision variable was used to define two distinct classes. Cases with a score below 3.00 were labelled low priority, while those with a score equal to or greater than 3.00 were considered high priority. This threshold of 3.00 was selected because it represents the midpoint of the decision scale, ensuring a clear separation between cases where the level of urgency is low and those that already demand closer attention. By doing so, the binary categorization avoids ambiguity in borderline cases and maximizes the contrast between the two groups. In contrast, the multi-class approach provided a more nuanced stratification by establishing four levels of priority. Specifically, cases with a decision score less than or equal to 1.49 were classified as not a priority; scores between 1.50 and 3.49 were labelled as low priority; scores between 3.50 and 4.49 were considered priority cases; and finally, scores equal to or greater than 4.50 were categorized as high priority. This classification schema enabled a more refined understanding of patient prioritization, supporting both coarse grained and fine grained decision modelling.

- Balanced Synthetic Data was performed to address the imbalance problem. In this context, we generate: (*i*) 52,525 synthetic instances were generated for class 0; and (*ii*) 50,000 synthetic instances were generated for class 1. After combining the real and synthetic data, both classes were balanced with 52,800 instances per class. In the case of multi-class, 24,964 synthetic data were generated for class 0, 24,253 data for class 1, 23,071 synthetic data for class 2 and 24,637 for class 3. In total there are 100,000 cases of which 96,925 are synthetic data. class 0 included 36 instances (scores less than or equal to 1.49), class 1 had 747 instances (scores ranging from 1.5 to 3.49), class 2 accounted for 1929 instances (scores ranging from 3.5 to 4.49), and class 3 comprised 363 instances (scores greater than or equal to 4.5). This process ensures that the final dataset was not only representative of the real world distributions but also provided equal representation for both classes, allowing the models to learn patterns effectively minority and majority outcomes.

Comparing the statistical distributions of real and synthetic data ensures that the synthetic data captures the patterns, variability, and relationships inherent in the original dataset [24].

To evaluate the similarity between real and synthetic datasets, a comprehensive set of descriptive statistical metrics was computed. These included mean, median, standard deviation, quartiles, kurtosis, and skewness [25]. Together, these measures allow to assess the central tendency, dispersion, distribution symmetry, and the presence of outliers, ensuring that the synthetic data faithfully reflects the statistical properties of the original dataset. The evaluation of these metrics is not only a validation step but also an integral part of the synthetic data generation process. In this study, the distributions of real and synthetic data were directly compared to ensure that the generated samples faithfully reproduced the statistical characteristics of the original dataset and did not introduce biases or unrealistic patterns..

These metrics were applied to compare the distributions of real and synthetic data at various stages of the generation process:

1. Initial Validation: The synthetic dataset was first evaluated using these statistical metrics to identify major deviations from real data. Any substantial differences in mean, standard deviation, or skewness signalled a need for parameter adjustments in the data generation model.
2. Iterative Refinement: Based on the initial findings, parameters of the SDV framework were fine tuned, ensuring that synthetic data maintained realistic variability and statistical consistency [16,24]. Multiple iterations of data generation and validation were conducted to minimize discrepancies.
3. Final Quality Check: Once the synthetic data closely aligned with the real dataset across all metrics, it was subjected to final validation by visualizing distributions (e.g., histograms, box plots) to confirm statistical similarity.

As part of the validation process, the methodology includes identifying any significant discrepancies between the distributions of real and synthetic data. When such differences are detected, the synthetic data generation process is fine tuned to better align with the statistical properties of the original dataset [24,26]. This iterative refinement ensures that ML models trained on synthetic data not only achieve high performance but also generalize effectively to real world scenarios [16,24,26].

4. Experiments and Results

To support the development of robust and generalizable ML models, synthetic data was generated to expand the training dataset, mitigate overfitting, and address the imbal-

ance across classes. Several generation strategies were explored, each aiming to refine the synthetic dataset in a way that preserved the statistical properties of the original data while improving class balance.

Fidelity metrics, including mean, standard deviation, skewness, and kurtosis, were computed separately for each variable in both the real and synthetic datasets. The absolute differences between corresponding metrics were then averaged across all variables to obtain a global fidelity score. This method allows for a fine-grained evaluation of the statistical similarity between real and synthetic data distributions.

4.1. Dataset

The dataset employed in this study comprises 2125 records that include both demographic information (age, gender, marital status, educational level, and employment status) and 20 healthcare decision-related variables (DEC1–DEC20). While the demographic variables provide contextual background concerning each respondent, the decision variables correspond to triage decision criteria typically used in healthcare environments to assess patient priority, urgency, and treatment allocation. Each decision captures how participants evaluate case severity based on predefined ethical and situational dimensions. Thus, the dataset does not contain direct clinical indicators such as diagnoses or laboratory data but rather focuses on healthcare triage decision making behaviour, which is a critical component of health system management and patient prioritization. This design ensures ethical compliance by avoiding sensitive personal data while maintaining strong healthcare relevance through the modelling of real decision processes.

The study relies on a purpose built dataset that captures a wide range of factors relevant to human decision making, personal attributes and moral values forming the empirical basis for our intelligent triage models. Data were obtained through an online survey completed by Portuguese participants. The questionnaire was administered between January 2020 and December 2023. Recruitment sought maximum heterogeneity: roughly 40 % were university students from multiple disciplines, while the remainder were adults reached via assorted public space approaches. Respondents averaged 24 years of age, the sample was 58 % female. The dataset encompasses data from 20 hypothetical rationing decision questions, each contrasting two individuals (Patient A vs. Patient B) who differ in personal or clinical characteristics. Participants, acting as decision makers, expressed their preference on a five point semantic differential scale: 1—definitely prioritize Patient A, 2—some priority to patient A, 3—no preference; 4—some preference for patient B, 5—definitely prioritize Patient B.

4.1.1. Questionnaire

The questionnaire design was based on previous work on healthcare rationing preferences and ethical decision making criteria, which employed 23 hypothetical scenarios [9,27,28]. In contrast, the present study streamlined the approach by using 20 carefully selected scenarios, each corresponding directly to one feature used in the model. Furthermore, while previous studies focused primarily on healthcare professionals as respondents, this work deliberately targeted lay citizens, aiming to capture public perspectives on medical prioritization.

Following Pinho & Araújo (2022), the six overarching rationing principles were disaggregated into nine specific criteria [9]:

- (i–ii) Clinical need: this criterion refers to the medical urgency of the patient, assessed through two components: Pain intensity represents the current physical suffering of the patient and signals a direct need for immediate intervention. Immediate risk of

death known as the Rule of Rescue, this sub-criterion prioritizes patients whose lives are in imminent risk of death.

- (iii–iv) Health maximization: this criterion seeks to maximize the health benefits resulting from treatment. It includes: Life expectancy gain measures the additional years of life that treatment is expected to provide. Quality of life improvement assesses how much the treatment can enhance the patient's daily well-being and functioning.
- (v) Social need: captured through Parenthood, this criterion considers that individuals with dependents (e.g., young children) may carry greater social responsibilities, which could justify prioritization.
- (vi) Age: this criterion evaluates the patient's age, often grounded in principles such as prioritarianism (giving preference to those who have lived fewer years) or the assumption of greater potential benefit in younger patients.
- (vii) Instrumental Value: refers to the essential or beneficial societal role of the patient (e.g., healthcare worker, firefighter). Prioritizing their recovery may result in broader social benefit.
- (viii) Negative Merit (risk taking lifestyle): considers whether the patient's condition resulted from voluntary risky behaviours (e.g., excessive alcohol, smoking, drug use). This criterion raises ethical debates around personal responsibility.
- (ix) Fair chance: operationalized through waiting time, it embodies the idea that everyone deserves a fair opportunity to receive treatment, regardless of clinical or social characteristics.

Table 1 lists the twenty decisions and maps to the corresponding criterion. In the raw file, each scenario response is stored in variables DEC1, DEC2, DEC3, DEC4, DEC5, DEC6, DEC7, DEC8, DEC9, DEC10, DEC11, DEC12, DEC13, DEC14, DEC15, DEC16, DEC17, DEC18, DEC19, DEC20, facilitating analysis of decision patterns (e.g., intuitive vs. deliberative choices). Additional columns record composite value orientations, conformity, benevolence, hedonism, openness to change, self-enhancement, supporting cluster analysis of participant profiles and group trends.

Table 1. Hypothetical Scenarios and Corresponding Rationing Criteria.

Decision	Patient A	Patient B	Criterion
1	Has a moderately painful disease	Has a very painful disease	Clinical need: Pain
2	Has been waiting 1 month	Has been waiting 6 months	Fair chance: Waiting time
3	Single, no dependants	Parent of three minors	Social need: Parenthood
4	Alcoholic with severe liver disease	Severe liver disease, never drank	Merit: Negative
5	Average citizen	Front-line physician	Instrumental value
6	20 % chance to live > 5 years	40 % chance to live > 5 years	Life expectancy
7	80 years old	40 years old	Age
8	Slight QoL improvement with treatment	Substantial QoL improvement	Quality of life
9	Joined queue today	Waiting 1 month	Fair chance

Table 1. *Cont.*

Decision	Patient A	Patient B	Criterion
10	Painful disease	Very painful disease	Clinical need: Pain
11	20 % chance to live > 5 years	80 % chance to live > 5 years	Life expectancy
12	Will die within 3 days w/o treatment	Will die immediately w/o treatment	Clinical need: Rule of Rescue
13	80 years old	20 years old	Age
14	Will die within 1 month w/o treatment	Will die within 1 week w/o treatment	Clinical need: Rule of Rescue
15	Slight QoL gain (poor → fair)	Large QoL gain (poor → very good)	Quality of life
16	HIV from unsafe sex	HIV from hospital transfusion	Negative Merit
17	25 years old	10 years old	Age
18	Average citizen	Scientist who discovered cures	Instrumental value
19	10 years old	Newborn	Age
20	Married, no children	Married, three school-age children	Social need

4.1.2. Descriptive Statistics

The structure of the dataset allows for a detailed examination of individual behaviours and value based preferences, supporting the identification of key patterns that shape triage decisions. By analyzing this wide range of variables, it becomes possible to gain deeper insight into how personal traits and ethical considerations influence decision making in healthcare scenarios.

Table 2 presents the descriptive statistics of the original dataset comprising 2125 triage-decision cases. The demographic variables indicate a heterogeneous participant profile: the mean age is 36.4 years (SD = 17.7), with a wide range (20–74 years), suggesting representation from both younger and older adults. The gender distribution (M = 0.63) shows a slight predominance of male participants, while marital status (mean = 1.73) and employment status (mean = 2.27) reveal moderate variability across social categories. The education level (mean = 4.77) indicates that most respondents have tertiary or postgraduate education.

Regarding the decision-related variables (DEC1–DEC20), the mean values range between 3.2 and 4.3, with medians concentrated around 4, indicating a general tendency toward moderate-to-high priority ratings in the triage decisions. The standard deviations, mostly between 0.9 and 1.1, reflect reasonable variability among participants' evaluations.

The negative skewness values observed in most decision variables (ranging approximately from -0.1 to -1.5) demonstrate a mild left skew, suggesting that participants more frequently assigned higher priority scores. This behaviour is consistent with the context of ethical triage, where individuals tend to favour caution or prioritize safety in uncertain clinical scenarios. The kurtosis values near zero indicate distributions close to normality, although a few variables (e.g., DEC1, DEC2, DEC9, DEC10) display mild leptokurtosis, implying a concentration of responses around higher values.

Table 2. Descriptive statistics of the original dataset.

Variable	Mean	Median	STD	Q1	Q3	Range	Kurtosis	Skewness
Age	36.43	31	17.66	20	50	74	−0.66	0.70
Gender	0.63	1	0.50	0.0	1.0	2.0	−1.43	−0.34
Marital Status	1.73	2.0	0.85	1.0	2.0	4.0	2.17	1.37
Education Level	4.77	4.0	1.59	4.0	6.0	8.0	0.17	0.0
Employment Status	2.27	2.0	1.41	1.0	3.0	5.0	−0.32	0.79
DEC1	4.16	4.0	0.98	4.0	5.0	4.0	1.99	−1.42
DEC2	4.26	5.0	0.99	4.0	5.0	4.0	2.13	−1.53
DEC3	3.86	4.0	0.94	3.0	5.0	4.0	0.32	−0.64
DEC4	3.76	4.0	0.94	3.0	5.0	4.0	−0.09	−0.48
DEC5	3.46	3.0	0.93	3.0	4.0	4.0	0.24	−0.19
DEC6	3.70	4.0	0.97	3.0	4.0	4.0	0.29	−0.59
DEC7	3.54	4.0	1.06	3.0	4.0	4.0	−0.11	−0.50
DEC8	3.78	4.0	0.97	3.0	5.0	4.0	0.14	−0.57
DEC9	4.16	4.0	0.97	4.0	5.0	4.0	1.60	−1.30
DEC10	4.11	4.0	0.92	4.0	5.0	4.0	1.32	−1.11
DEC11	3.91	4.0	1.01	3.0	5.0	4.0	0.42	−0.82
DEC12	3.93	4.0	1.08	3.0	5.0	4.0	0.14	−0.85
DEC13	3.57	4.0	1.13	3.0	4.0	4.0	−0.36	−0.49
DEC14	3.85	4.0	1.06	3.0	5.0	4.0	0.20	−0.79
DEC15	3.83	4.0	1.00	3.0	5.0	4.0	0.07	−0.63
DEC16	3.73	4.0	0.98	3.0	5.0	4.0	−0.23	−0.32
DEC17	3.58	4.0	1.03	3.0	4.0	4.0	−0.20	−0.35
DEC18	3.33	3.0	0.93	3.0	4.0	4.0	0.46	−0.11
DEC19	3.24	3.0	1.09	3.0	4.0	4.0	−0.32	−0.14
DEC20	3.63	4.0	0.93	3.0	4.0	4.0	0.06	−0.27

If we divide the current dataset by different classes there were 36 cases with a value less than or equal to 1.49 for class 0 (not a priority), 747 cases with values between 1.50 and 3.49 for class 1 (low priority cases), 1929 for class 2 (Priority cases) with values between 3.50 and 4.49 and 363 cases with a value greater or equal to 4.50 for class 3 (High priority cases). This enriched dataset is then used to train ML models, with feature selection methods helping to pinpoint the most relevant factors for determining patient priority. These models are designed with interpretability in mind, ensuring that their results are both understandable and practically useful in real world clinical settings. In doing so, the dataset plays a central role in answering the study's research questions and in demonstrating the validity of the proposed approach.

4.2. Pre-Processing

The original dataset consisted of 2125 cases and presented a clear class imbalance, both in the binary and multi-class classification scenarios. In the binary configuration, class 0 included only 275 instances corresponding to responses rated between 0 and 3 while class 1 comprised 2800 instances, associated with ratings higher that 3. It is important to note that the dataset does not present a consistent total number of valid responses per case, as the overall average score used to classify each instance could not always be calculated. This was due to missing values (*NaN*) in the decision-related questions, which led to the exclusion of incomplete cases from the classification process.

In the multi-class setup, the distribution was similarly skewed: class 0 included 36 instances (scores less than or equal to 1.49), class 1 had 747 instances (scores ranging from 1.5 to 3.49), class 2 accounted for 1929 instances (scores ranging from 3.5 to 4.49), and class 3 comprised 363 instances (scores greater than or equal to 4.5).

4.3. Binary Scenario

Refined Synthetic Data. The final approach focused on precisely addressing the inherent class imbalance by controlling the number of synthetic instances generated per class. Unlike previous strategies, this method did not simply replicate the global distribution of the original dataset or apply balancing as a preliminary step. Instead, it adopted a targeted

strategy, determining the exact number of synthetic cases needed for each class so that, when combined with the original real data, the final dataset would total 100,000 instances with fully balanced class representation.

This refined approach was applied to both the binary and multi-class classification settings, as it proved to be the most effective and robust strategy for producing high quality synthetic data. Its goal was not only to correct imbalances but also to provide a strong, generalizable foundation for training ML models with fair exposure to all class scenarios.

Data Distribution. A refined synthetic data generation strategy was implemented that explicitly considered class balance while still being based on the original imbalanced real data, which was applied to the binary and multi-class case. The goal was to ensure that the generated dataset reflected the real world distribution while also providing a sufficient number of cases for each class to train the models effectively. A crucial aspect of validating the synthetic data is comparing its distribution with that of the real data. To ensure that the generated data reflects the statistical characteristics of the real dataset, a thorough comparative analysis was conducted on each variable and several key variables were compared using statistical metrics.

Table 3 compares the distribution of all variables between the original and synthetic datasets for the third approach. In this strategy, different amounts of synthetic data were generated per class to counteract the real dataset's imbalance, ensuring that after combining both sources, each class was equally represented.

Table 3. Comparison of Real and Synthetic Data of the demographic variables for the refined synthetic data in binary scenario.

Statistic	Age		Gender		Marital Status		Education Level		Employment Status	
	Real	Synth	Real	Synth	Real	Synth	Real	Synth	Real	Synth
Mean	36.43	38.05	0.63	0.60	1.73	1.73	4.77	4.78	2.27	2.25
Median	31.0	35.0	1.0	1.0	2.0	2.0	4.0	4.0	2.0	2.0
STD	17.67	19.31	0.50	0.51	0.85	0.84	1.59	1.63	1.41	1.41
Q1	20.0	24.0	0.0	0.0	1.0	1.0	4.0	4.0	1.0	1.0
Q3	50.0	57.0	1.0	1.0	2.0	2.0	6.0	6.0	3.0	3.0
Range	74.0	73.0	2.0	2.0	4.0	4.0	8.0	8.0	5.0	5.0
Kurtosis	−0.66	−0.63	−1.43	−1.40	2.17	2.27	0.17	0.08	−0.32	−0.36
Skewness	0.70	0.71	−0.34	−0.18	1.37	1.40	−0.00	−0.06	0.77	0.78

Table 3 compares real and synthetic data under the refined synthetic data approach, showing high alignment across most variables. This method successfully maintains the structural integrity of all variables while introducing a controlled level of variability that strengthens the dataset's robustness. For instance, the age distribution in the synthetic data, although slightly shifted towards older individuals, broadens the variability without distorting the overall population profile. This is particularly valuable as it ensures representation of wider patient scenarios, enriching model generalization.

Similarly, gender, marital status, and employment status show very consistency between real and synthetic data, with only marginal statistical deviations that fall within acceptable tolerance. The education level variable also exhibits nearly perfect alignment, proving that categorical structures are reliably replicated. The minor differences observed they reduce the risk of overfitting to narrow real world patterns, while preserving the realism needed for credible decision support models. Ranges remain consistent, and categorical levels are fully preserved, ensuring no loss of semantic meaning or data integrity.

In summary, the synthetic dataset successfully replicates the key statistical characteristics of the original data while enhancing balance and diversity.

Table 4 compares the distributions of the decision related variables (DEC1 to DEC5) between the real dataset and the synthetic data generated under the refined synthetic data. This approach focuses on generating synthetic samples in a way that, when combined with

the original data, results in balanced class representation across the target variable while maintaining distributional characteristics of the predictors.

Table 4. Comparison of Real and Synthetic Data for Decision Variables (DEC1–DEC5) in binary scenario.

Statistic	DEC1		DEC2		DEC3		DEC4		DEC5	
	Real	Synth	Real	Synth	Real	Synth	Real	Synth	Real	Synth
Mean	4.16	3.85	4.26	3.60	3.86	3.40	3.76	3.29	3.46	3.15
Median	4.0	4.0	5.0	4.0	4.0	4.0	4.0	4.0	3.0	3.0
STD	0.98	1.29	0.99	1.37	0.94	1.24	0.99	1.32	0.93	1.16
Q1	4.0	3.0	4.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0
Q3	5.0	5.0	5.0	5.0	5.0	4.0	5.0	4.0	4.0	4.0
Range	4.0	5.0	4.0	5.0	4.0	5.0	4.0	5.0	4.0	5.0
Kurtosis	1.99	0.57	2.13	−0.02	0.32	0.31	−0.09	−0.27	0.24	0.12
Skewness	−1.42	−1.05	−1.53	−0.87	−0.64	−0.88	−0.4800	−0.66	−0.19	−0.54

The statistical comparison across the five decision variables indicates that the synthetic data captures key distributional characteristics present in the real dataset. For all variables, the medians and inter-quartile ranges are preserved, reflecting consistency in the core distribution shape. However, the synthetic data exhibits slightly lower means across all variables, coupled with higher standard deviations. This indicates a broader spread and a modest shift in central tendency, potentially reflecting the increased representation of lower values in the synthetic samples. Despite these variations, the alignment in median and quartile values supports the conclusion that the synthetic data remains representative of the real data's structure. This validates the suitability of the generated samples for downstream tasks, while fulfilling the goal of class balancing imposed by the refined synthetic data.

Data Balancing. The method began by analyzing the number of real cases in each class (275 in class 0 and 2800 in class 1). Based on this, the appropriate number of synthetic samples was individually generated for each class to ensure that both reached an equal representation in the final dataset. This means generating proportionally more data for the underrepresented class (class 0), and less for the dominant class (class 1), correcting the existing bias without distorting the statistical properties of the original data.

A total of 49,725 synthetic instances were generated for class 0 and 47,200 for class 1. When combined with the original real data, each class reached exactly 50,000 instances, resulting in a perfectly balanced final dataset. This equal class distribution provides a solid foundation for developing classification models, enabling the algorithm to learn how to predict patient priority levels without being affected by the structural bias present in the original dataset. In addition to improving class balance, this controlled generation ensured that each synthetic record adhered to the learned statistical relationships and logical consistency of the original dataset. By doing so, the model avoided common pitfalls associated with naive oversampling, such as redundancy or data artefacts.

4.4. Multi-Class Scenario

Refined Synthetic Data. Following the success of the refined balancing strategy in the binary classification scenario, the same methodology was extended to the multi-class configuration. Given the more complex distribution of classes in this scenario, with varying levels of under-representation, a careful and systematic approach was adopted to ensure an even class distribution while preserving the statistical integrity of the original data. Rather than applying a generic oversampling method or maintaining the natural imbalance, this refined approach aimed to generate the exact number of synthetic instances needed per class. The objective was to ensure that, when combined with the real dataset, each of the four classes would be equally represented, resulting in a final dataset composed of 100,000 balanced instances, 25,000 per class.

Data Distribution. This scenario involved four distinct priority levels: non-priority, low priority, priority, and high priority. The synthetic data was carefully generated to ensure

equal representation across these categories, with each class totalling 25,000 instances after combining real and synthetic records. Before moving forward with the evaluation of individual variables, it was essential to validate the overall distribution of the synthetic data. A comparative distribution analysis was performed between the real and synthetic data across all variables, highlighting how well the synthetic data replicates the structure of the real dataset in this multi-class setting. This analysis served as a critical step in verifying the fidelity and reliability of the synthetic dataset. Particular attention was given to variables that influence classification performance, ensuring that the synthetic data maintained both statistical consistency and semantic realism.

Table 5 presents a comparison of the distribution for the variables between the real dataset and the synthetic data generated under the final multi-class balancing approach. The table provides an overview of key distributional statistics to assess the fidelity of the synthetic data in replicating the structure of the original dataset.

Table 5. Comparison of Real and Synthetic Data of demographic variables for the Refined Synthetic Data in multi-class scenario.

Statistic	Age		Gender		Marital Status		Education Level		Employment Status	
	Real	Synth	Real	Synth	Real	Synth	Real	Synth	Real	Synth
Mean	36.43	37.12	0.63	0.61	1.73	1.73	4.77	4.78	2.27	2.29
Median	31.0	32.0	1.0	1.0	2.0	2.0	4.0	4.0	2.0	2.0
STD	17.66	15.57	0.50	0.50	0.85	0.82	1.60	1.61	1.41	1.43
Q1	20.0	25.0	0.0	0.0	1.0	1.0	4.0	4.0	1.0	1.0
Q3	50.0	46.0	1.0	1.0	2.0	2.0	6.0	6.0	3.0	3.0
Range	74.0	75.0	2.0	2.0	4.0	4.0	8.0	8.0	5.0	5.0
Kurtosis	−0.66	0.47	−1.43	−1.46	2.17	2.28	0.17	0.14	−0.32	−0.45
Skewness	0.70	1.04	−0.34	−0.26	1.37	1.34	−0.00	−0.05	0.77	0.76

The refined synthetic dataset under the multi-class scenario shows strong alignment with the real data while introducing small variations that enhance representativeness. The age variable presents a slightly higher mean and skewness, capturing a broader spread of cases and improving balance across groups.

Gender, marital status, and education level remain almost identical to the real data, preserving structural fidelity. For employment status, the synthetic data introduces slightly higher variability, which is beneficial for generalization. Overall, this refined approach ensures both realism and controlled heterogeneity, making it especially suited for robust multi-class modelling.

Table 6 presents a comparative summary of the statistical distribution of five decision related variables (DEC1–DEC5) between the real dataset and the synthetically generated dataset, both derived under the final multi-class balancing strategy. This evaluation is crucial to assess whether the synthetic data can reliably mimic the key distributional characteristics of decision making variables found in the original dataset, which are central to downstream modelling and interpretation.

Table 6. Comparison of Real and Synthetic Data for Decision Variables (DEC1–DEC5) in multi-class scenario.

Statistic	DEC1		DEC2		DEC3		DEC4		DEC5	
	Real	Synth	Real	Synth	Real	Synth	Real	Synth	Real	Synth
Mean	4.16	3.73	4.26	3.78	3.86	3.54	3.76	3.39	3.46	3.15
Median	4.0	4.0	5.0	4.0	4.0	4.0	4.0	4.0	3.0	3.0
STD	0.98	1.69	0.99	1.66	0.94	1.62	0.99	1.64	0.93	1.46
Q1	4.0	3.0	4.0	3.0	3.0	2.0	3.0	2.0	3.0	2.0
Q3	5.0	5.0	5.0	5.0	5.0	5.0	5.0	5.0	4.0	4.0
Range	4.0	5.0	4.0	5.0	4.0	5.0	4.0	5.0	4.0	5.0
Kurtosis	1.99	−0.33	2.13	−0.44	0.32	−0.60	−0.09	−0.82	0.24	−0.47
Skewness	−1.42	−0.99	−1.53	−0.92	−0.64	−0.76	−0.48	−0.63	−0.19	−0.56

The descriptive statistics demonstrate that the synthetic data maintains a close approximation to the real dataset across all five decision variables. The mean and median values

are generally consistent, with DEC1 and DEC2 showing slightly higher central tendency in the real data, while the synthetic data exhibits marginally increased dispersion as reflected in higher standard deviations for all variables.

The general shape and symmetry of the distributions are preserved, the synthetic dataset offers a faithful representation of the real decision variables, supporting its adequacy for subsequent analytical tasks within the multi-class framework.

The statistical comparison confirmed that the synthetic data is a highly reliable representation of the real dataset. The preservation of means, medians, and quartiles of the variables highlights the model's ability to faithfully replicate central tendency and variability, ensuring that the synthetic data reflects the essential statistical structure of the real dataset. While minor differences exist they do not significantly impact the overall validity of the dataset. The high degree of similarity in statistical metrics suggests that the synthetic data can be confidently used for analysis and modelling, maintaining the essential characteristics of the original real world data.

Data Balancing. The process began with an analysis of the real data distribution across the four priority levels: class 0 (non-priority), class 1 (low priority), class 2 (priority), and class 3 (high priority). Based on the number of real instances in each class, synthetic data was generated in a targeted manner to complement the deficit in each category. For class 0 has been generated 24,955 data samples, 24,253 synthetic data for class 1, 23,071 for class 2, and 24,637 synthetic data for class 3. When combined with the respective real data, these additions brought each class to exactly 25,000 instances. This resulted in a fully balanced multi-class dataset with a total of 100,000 examples, ensuring fair and equal representation of all priority levels for model training.

In addition to achieving balance, special care was taken to ensure that the synthetic records remained statistically and semantically coherent. The average values of key decision variables within each class stayed within the expected ranges defined for their respective categories. Variables such as age, political orientation, and service usage patterns remained within realistic and medically plausible boundaries. By applying this refined and controlled synthesis approach, the resulting multi-class dataset provides a robust and fair foundation for machine learning applications, enhancing the model's ability to learn from all classes equally and generalize effectively across different patient profiles.

Given this overall performance, this approach was selected as the basis for developing the algorithm integrated into the patient triage application, as it proved to be the most effective in terms of distributional fidelity and data balancing. The next step will be to apply machine learning algorithms to this dataset to build and validate the predictive components of the system.

4.5. Comparative Analysis with Related Work

Several recent studies have reported promising outcomes using generative models such as CTGAN, CopulaGAN, and VAE for synthetic health data generation. According to Hernandez et al. (2022) review [11], the heathcare domain concentrates the generation of synthetic data using GAN. Although neural-based models such as GANs and VAEs have demonstrated strong capabilities in generating complex synthetic data, their application in healthcare remains problematic. These architectures frequently operate with limited interpretability, which complicates the validation of outcomes and raises concerns regarding trust and accountability in clinical contexts. Moreover, they tend to suffer from training instabilities, including convergence failures and overfitting to dominant patterns, which may suppress rare but medically relevant cases. Their performance is also highly dependent on the correct selection of hyperparameters and hardware optimization, resulting in increased

computational demands and limited scalability when applied to large or heterogeneous healthcare datasets.

Table 7 summarizes key characteristics of recent studies on synthetic data generation in healthcare and compares them with the proposed approach. Previous works, such as those by Nasimov et al. (2024) [3] and Abedi et al. (2022) [8], primarily relied on GAN-based architectures and focused on binary classification problems with relatively limited dataset sizes. These models, although effective in small-scale experiments, often suffer from training instability and limited generalization.

In contrast, the proposed SDV-based approach supports both binary and multiclass scenarios and demonstrates superior scalability, expanding the dataset from 2125 real samples to 100,000 synthetic observations while maintaining statistical fidelity. This scalability, combined with improved transparency and lower computational complexity, makes the proposed method a robust and adaptable alternative for healthcare data synthesis compared with traditional GAN-based techniques.

Table 7. Comparison of Related Work and the Proposed Approach.

Approach	Method	Classification Scope	Domain	Original Sample Size	Final Data Sample
[3]	GAN	Binary	Health	766	10,000
[7]	GAN	Not Available	Health	Not Available	Not Available
[8]	GAN	Binary	Health	569	Not Available
Proposed	SVD	Binary and Multi-class	Health	3085	100,000

4.6. Discussion

The experimental results demonstrate that the proposed synthetic data generation approach, based on the SDV Gaussian Copula Synthesizer, effectively reproduces the statistical structure of the original healthcare dataset while ensuring data privacy. The analysis of distributional metrics, including mean, standard deviation, skewness, and kurtosis, confirms a high degree of similarity between real and synthetic data, indicating that the generated samples maintain the integrity and variability of the original variables. These findings suggest that the probabilistic, statistically grounded nature of SDV provides a stable and explainable alternative to neural generative models such as GANs and VAEs, which often suffer from training instability and overfitting.

When compared with previous studies (Table 7), the proposed approach extends existing work in several dimensions. Unlike prior GAN-based studies that focused primarily on binary classification and small datasets [3,8], this study successfully handles both binary and multiclass healthcare triage problems. Moreover, the SDV framework enables the expansion of the dataset from 2125 real to 100,000 synthetic records without compromising statistical fidelity. This scalability highlights the robustness of the probabilistic copula-based modelling approach and its ability to support applications that require larger and more diverse datasets.

From a practical view, these results have important implications for data-driven decision making in healthcare. The ability to generate realistic, privacy-preserving synthetic data enables hospitals, research institutions, and policymakers to conduct advanced modelling and triage simulations without exposing sensitive patient information. This can facilitate the development of predictive tools, improve ethical resource allocation, and support compliance with data protection regulations.

However, several limitations should be acknowledged. First, the present study focused on statistical fidelity and did not include a full evaluation of model performance metrics or statistical distance measures such as Jensen–Shannon or Kolmogorov–Smirnov divergence. Second, the dataset used contained primarily demographic and decision-related attributes, which may not fully represent the complexity of clinical data. Furthermore, the next stage of this research will focus on developing an explainable prioritization platform that leverages the generated synthetic data to support transparent and interpretable triage decision making. This platform will combine data-driven prioritization models with explainability mechanisms, enabling healthcare professionals to better understand and justify the recommendations produced by AI-assisted systems.

In summary, the proposed SDV-based approach demonstrates that statistically grounded models can generate scalable, high-fidelity, and privacy-respecting synthetic data suitable for healthcare triage applications. Ongoing work will extend this research to include more complex data modalities, quantitative similarity metrics, and open access to the implementation upon request, thus reinforcing transparency and reproducibility in synthetic data generation research.

From a privacy perspective, the generation of synthetic data inherently supports data protection principles. Because the SDV Gaussian Copula Synthesizer reproduces only the statistical relationships among variables and not the actual records, no individual information from the original dataset is retained. This mechanism effectively prevents data re-identification, enabling the use of realistic datasets for analysis and model training in compliance with privacy regulations.

5. Conclusions

The analysis demonstrates that synthetic data generation constitutes an effective strategy for addressing the limitations of the original dataset, particularly the severe class imbalance and the relatively small sample size in minority groups. Across all approaches, the comparative analysis confirmed that the generated data accurately preserved the statistical properties of the real dataset, both for sociodemographic variables and decision-related attributes. Central tendency, dispersion, and higher-order metrics such as skewness and kurtosis were consistently aligned, ensuring that the synthetic records retained the shape and structure of the original distributions without introducing distortions.

Among the three strategies tested, the refined synthetic data approach proved to be the most robust and reliable. By explicitly controlling the number of synthetic instances generated for each class, this method ensured full class balance in both binary and multi-class classification scenarios, while simultaneously maintaining distributional fidelity across key predictors. This not only mitigates the risk of overfitting but also provides a more generalizable and fair basis for machine learning model training.

Overall, the results highlight the dual role of synthetic data: on the one hand, it preserves the realism and statistical integrity of the source dataset; on the other, it enriches the training environment by correcting imbalances and broadening variability. This positions the refined synthetic dataset as a solid foundation for the development of interpretable and trustworthy ML models in medical triage, ultimately supporting more equitable and data-driven decision making in healthcare resource allocation.

While the results of this study confirm the potential of the SDV Gaussian Copula Synthesizer to generate high-fidelity and privacy-preserving synthetic data for healthcare applications, several limitations must be recognized. The present work focused primarily on statistical validation, without incorporating full performance or distance-based similarity metrics. Moreover, the dataset used contained mostly demographic and decision-related variables, which limits the scope of inference for clinical contexts. Future work will therefore

expand the analysis by integrating clinical indicators and evaluating additional fidelity and privacy metrics. Building upon these foundations, the next stage of this research will involve the development of an explainable prioritization platform that leverages the generated synthetic data to support transparent, interpretable, and ethically aligned triage decisions in real-world healthcare environments.

Author Contributions: Conceptualization, C.G., F.L. and M.P.; methodology, C.G., F.L. and M.P.; software, C.G. and F.L.; validation, C.G., F.L. and M.P.; formal analysis, C.G., F.L. and M.P.; investigation, C.G., F.L. and M.P. resources, F.L. and M.P.; data curation, C.G. and F.L. ; writing—original draft preparation, C.G., F.L. and M.P.; writing—review and editing, F.L. and M.P.; visualization, C.G., F.L. and M.P.; supervision, F.L. and M.P.; project administration, F.L. and M.P.; funding acquisition, F.L. and M.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., under the Programme Contract UID/05105/2025.

Institutional Review Board Statement: This manuscript does not contain clinical studies or patient-level data. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee (DL. 80/2018 de 15 outubro; Regulamento (UE) 2016/679 do Parlamento Europeu e do Conselho de 27 de abril de 2016; Regulamento da CESUPT de 2 de junho de 2020) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Statement: Written informed consent was obtained from all participants before being included in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ML	Machine Learning
SVD	Synthetic Data Vault
VAE	Variational Autoencoders
GAN	Generative Adversarial Networks
CTGAN	Conditional Tabular Generative Adversarial
FPCA	Functional Principal Component Analysis
ECDF	Empirical Cumulative Distribution Functions
GMM	Gaussian Mixture Models
CSV	Comma-separated values
BN	Bayesian-Network

References

1. Pezoulas, V.C.; Zaridis, D.I.; Mylona, E.; Androutsos, C.; Apostolidis, K.; Tachos, N.S.; Fotiadis, D.I. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Comput. Struct. Biotechnol. J.* **2024**, *23*, 2892–2910. https://doi.org/10.1016/j.csbj.2024.07.005.
2. Liu, K.; Altman, R.B. Conditional Generative Models for Synthetic Tabular Data: Applications for Precision Medicine and Diverse Representations. *Annu. Rev. Biomed. Data Sci.* **2025**, *8*, 21–49. https://doi.org/10.1146/annurev-biodatasci-103123-094844.
3. Nasimov, R.; Nasimova, N.; Mirzakhalilov, S.; Tokdemir, G.; Rizwan, M.; Abdusalomov, A.; Cho, Y.I. GAN-Based Novel Approach for Generating Synthetic Medical Tabular Data. *Bioengineering* **2024**, *11*, 1288. https://doi.org/10.3390/bioengineering11121288.
4. Endres, M.; Mannarapotta Venugopal, A.; Tran, T.S. Synthetic Data Generation: A Comparative Study. In Proceedings of the IDEAS '22: Proceedings of the 26th International Database Engineered Applications Symposium, New York, NY, USA, 22–24 August 2022. https://doi.org/10.1145/3548785.3548793.

5. Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. Modeling tabular data using conditional GAN. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Curran Associates Inc.: Red Hook, NY, USA, 2019.
6. Jutras-Dubé, P.; Al-Khasawneh, M.B.; Yang, Z.; Bas, J.; Bastin, F.; Cirillo, C. Copula-based transferable models for synthetic population generation. *Transp. Res. Part C: Emerg. Technol.* **2024**, *169*, 104830. https://doi.org/10.1016/j.trc.2024.104830.
7. Hahn, W.; Schütte, K.; Schultz, K.; Wolkenhauer, O.; Sedlmayr, M.; Schuler, U.; Eichler, M.; Bej, S.; Wolfien, M. Contribution of Synthetic Data Generation towards an Improved Patient Stratification in Palliative Care. *J. Pers. Med.* **2022**, *12*, 1278. https://doi.org/10.3390/jpm12081278.
8. Abedi, M.; Hempel, L.; Sadeghi, S.; Kirsten, T. GAN-Based Approaches for Generating Structured Data in the Medical Domain. *Appl. Sci.* **2022**, *12*, 7075. https://doi.org/10.3390/app12147075.
9. Pinho, M.; Araújo, A. How to fairly allocate scarce medical resources? Controversial preferences of healthcare professionals with different personal characteristics. *Health Econ. Policy Law* **2022**, *17*, 398–415. https://doi.org/10.1017/S1744133121000190.
10. Sun, Y.; Cuesta-Infante, A.; Veeramachaneni, K. Learning vine copula models for synthetic data generation. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; AAAI Press: Washington, DC, USA, 2019; AAAI'19/IAAI'19/EAAI'19. https://doi.org/10.1609/aaai.v33i01.33015049.
11. Hernández, M.; Epelde, G.; Alberdi, A.; Cilla, R.; Rankin, D. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing* **2022**, *510*, 235–257. https://doi.org/10.1016/j.neucom.2022.04.053.
12. Meyer, D.; Nagler, T. Synthia: Multidimensional synthetic data generation in Python. *J. Open Source Softw.* **2021**, *6*, 2863. https://doi.org/10.21105/joss.02863.
13. de Arriba-Perez, F.; Garcia-Mendez, S.; Leal, F.; Malheiro, B.; Burguillo-Rial, J.C. Balancing Plug-In for Stream-Based Classification. In *Information Systems and Technologies: WorldCIST 2023, Volume 1*; Álvaro, R., Adeli, H., Dzemyda, G., Moreira, F., Colla, V., Eds.; Springer: Cham, Switzerland, 2024; pp. 65–74. https://doi.org/10.1007/978-3-031-45642-8_6.
14. Nelsen, R.B. *An Introduction to Copulas*, 2nd ed.; Springer: New York, NY, USA, 2006.
15. Nikoloulopoulos, A.K. Efficient and feasible inference for high-dimensional normal copula regression models. *Comput. Stat. Data Anal.* **2023**, *179*, 107654. https://doi.org/https://doi.org/10.1016/j.csda.2022.107654.
16. Patki, N.; Wedge, R.; Veeramachaneni, K. The Synthetic Data Vault. In Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, Canada, 17–19 October 2016; pp. 1–10. https://doi.org/10.1109/DSAA.2016.49.
17. Jadon, A.; Kumar, S. Leveraging Generative AI Models for Synthetic Data Generation in Healthcare: Balancing Research and Privacy. In Proceedings of the 2023 International Conference on Smart Applications, Communications and Networking (SmartNets), Istanbul, Turkey, 25–27 July 2023; pp. 1–4. https://doi.org/10.1109/SmartNets58706.2023.10215825.
18. Khorshidi, H.A.; Aickelin, U. A Synthetic Over-sampling method with Minority and Majority classes for imbalance problems. *Knowl. Inf. Syst.* **2025**, *67*, 5965–5998. https://doi.org/10.1007/s10115-025-02394-6.
19. Torfi, A.; Fox, E.A.; Reddy, C.K. Differentially private synthetic medical data generation using convolutional GANs. *Inf. Sci.* **2022**, *586*, 485–500. https://doi.org/https://doi.org/10.1016/j.ins.2021.12.018.
20. Gurcan, F.; Soylu, A. Learning from Imbalanced Data: Integration of Advanced Resampling Techniques and Machine Learning Models for Enhanced Cancer Diagnosis and Prognosis. *Cancers* **2024**, *16*, 3417. https://doi.org/10.3390/cancers16193417.
21. Salmi, M.; Atif, D.; Oliva, D.; Abraham, A.; Ventura, S. Handling imbalanced medical datasets: Review of a decade of research. *Artif. Intell. Rev.* **2024**, *57*, 273. https://doi.org/10.1007/s10462-024-10884-2.
22. Rezvani, S.; Wang, X. A broad review on class imbalance learning techniques. *Appl. Soft Comput.* **2023**, *143*, 110415. https://doi.org/10.1016/j.asoc.2023.110415.
23. Abdulsadig, R.S.; Villegas, E. A comparative study in class imbalance mitigation when working with healthcare data. *Front. Digit. Health* **2024**, *6*, 1377165. https://doi.org/https://doi.org/10.3389/fdgth.2024.1377165.
24. Yan, C.; Zhang, Z.; Nyemba, S.; Li, Z. Generating Synthetic Electronic Health Record Data Using Generative Adversarial Networks: Tutorial. *JMIR AI* **2024**, *3*, e52615. https://doi.org/10.2196/52615.
25. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag. Int. J.* **2009**, *45*, 427–437. https://doi.org/https://doi.org/10.1016/j.ipm.2009.03.002.
26. Ayyanar, M.; Jegananathan, S.; Parthasarathy, S.; Jayaraman, V.; Lakshminarayanan, A.R. Predicting the Cardiac Diseases using SelectKBest Method Equipped Light Gradient Boosting Machine. In Proceedings of the 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 25–27 May 2022. https://doi.org/10.1109/ICOEI53556.2022.9777224.

27. Kasemsup, V.; Schommer, J.; Cline, R.; Hadsall, R. Citizen's preferences regarding principles to guide health care allocation decisions in Thailand. *Value Health* **2008**, *11*, 1194–1202. https://doi.org/10.1111/j.1524-4733.2008.00321.x.
28. Pinho, M.; Veiga, P. Attitudes of health professionals concerning bedside rationing criteria: A survey from Portugal. *Health Econ. Policy Law* **2020**, *15*, 113–127. https://doi.org/10.1017/S1744133118000403.