



# A computational pipeline for data augmentation towards the improvement of disease classification and risk stratification models: A case study in two clinical domains

Vasileios C. Pezoulas<sup>a</sup>, Grigoris I. Grigoriadis<sup>a</sup>, George Gkois<sup>a</sup>, Nikolaos S. Tachos<sup>a</sup>, Tim Smole<sup>b</sup>, Zoran Bosnić<sup>b</sup>, Matej Pičulin<sup>b</sup>, Iacopo Olivotto<sup>c</sup>, Fausto Barlocco<sup>c</sup>, Marko Robnik-Šikonja<sup>b</sup>, Djordje G. Jakovljević<sup>d</sup>, Andreas Goules<sup>e</sup>, Athanasios G. Tzioufas<sup>e</sup>, Dimitrios I. Fotiadis<sup>a,f,\*</sup>

<sup>a</sup> Unit of Medical Technology and Intelligent Information Systems, Department of Materials Science and Engineering, University of Ioannina, Ioannina, GR45110, Greece

<sup>b</sup> Faculty of Computer and Information Science, University of Ljubljana, Večna Pot 113, 1000, Ljubljana, Slovenia

<sup>c</sup> Department of Experimental and Clinical Medicine, University of Florence and Cardiomyopathies Unit, Azienda Ospedaliera Careggi, Florence, Italy

<sup>d</sup> Faculty of Medical Sciences, Newcastle University, Newcastle Upon Tyne, UK and with the Faculty of Health and Life Sciences, Coventry University, Coventry, UK

<sup>e</sup> Department of Pathophysiology, Faculty of Medicine, National and Kapodistrian University of Athens (NKUA), GR 15772, Athens, Greece

<sup>f</sup> Department of Biomedical Research, FORTH-IMBB, Ioannina, GR45110, Greece

## ARTICLE INFO

### Keywords:

Artificial intelligence  
Data augmentation  
Virtual population generation  
Lymphoma classification  
HCM risk stratification

## ABSTRACT

Virtual population generation is an emerging field in data science with numerous applications in healthcare towards the augmentation of clinical research databases with significant lack of population size. However, the impact of data augmentation on the development of AI (artificial intelligence) models to address clinical unmet needs has not yet been investigated. In this work, we assess whether the aggregation of real with virtual patient data can improve the performance of the existing risk stratification and disease classification models in two rare clinical domains, namely the primary Sjögren's Syndrome (pSS) and the hypertrophic cardiomyopathy (HCM), for the first time in the literature. To do so, multivariate approaches, such as, the multivariate normal distribution (MVND), and straightforward ones, such as, the Bayesian networks, the artificial neural networks (ANNs), and the tree ensembles are compared against their performance towards the generation of high-quality virtual data. Both boosting and bagging algorithms, such as, the Gradient boosting trees (XGBoost), the AdaBoost and the Random Forests (RFs) were trained on the augmented data to evaluate the performance improvement for lymphoma classification and HCM risk stratification. Our results revealed the favorable performance of the tree ensemble generators, in both domains, yielding virtual data with goodness-of-fit 0.021 and KL-divergence 0.029 in pSS and 0.029, 0.027 in HCM, respectively. The application of the XGBoost on the augmented data revealed an increase by 10.9% in accuracy, 10.7% in sensitivity, 11.5% in specificity for lymphoma classification and 16.1% in accuracy, 16.9% in sensitivity, 13.7% in specificity in HCM risk stratification.

## 1. Introduction

The current advances in data science have led to the development of an emerging branch of applications which focuses on the augmentation of medical data. Its aim is to shed light into the underlying structure of clinical problems towards the development of robust machine learning models for predicting disease outcomes and their risk levels. Virtual

population generation [1] refers to the development of computational methods that can be used to generate artificial (or synthetic) patient data by producing virtual distributions like those in the real world. Its desired usage is to enhance the statistical power of clinical research databases with significant lack of population size. Data augmentation [2] refers to the aggregation of the real with the virtual patient data to yield AI (artificial intelligence) models with increased performance for

\* Corresponding author. Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, University of Ioannina, GR45110, Ioannina, Greece.

E-mail address: [fotiadis@uoi.gr](mailto:fotiadis@uoi.gr) (D.I. Fotiadis).

<https://doi.org/10.1016/j.combiomed.2021.104520>

Received 8 March 2021; Received in revised form 13 May 2021; Accepted 24 May 2021

Available online 6 June 2021

0010-4825/© 2021 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

classification tasks. For example, in medical imaging, data augmentation refers to the application of data mirroring and data cropping methods to enhance the performance of the existing deep learning models for image segmentation by increasing the size of the input training data with virtual training data. Clinical data augmentation refers to the aggregation of the real with the high-quality virtual clinical data to address various clinical unmet needs including the development of robust risk stratification and disease classification models, as well as, the detection of biomarkers, among others.

As a result, the clinical value of data augmentation lies on the quality of the virtually generated data. Indeed, the aggregation of the real data with poor quality virtual data (i.e., virtual data with increased divergence and reduced similarity with the real data) is expected to have a negative impact on the performance of the AI models. As a matter of fact, particular emphasis must be given on the development of robust virtual population generators. In addition, prior the development of the virtual population generators it is crucial to apply data curation methods towards the detection and removal of data recording errors, inconsistent data types and problematic fields that are present in the input clinical data. This step is important since the application of the virtual population generators on contaminated data might produce virtual data with poor performance and reduced statistical power of downstream applications. Thus, data curation must be applied to meet standard data quality criteria in terms of data completeness and conformity [9–11], among others.

The state-of-the-art methods for virtual population generation can be classified into two major categories; the parametric methods which resample instances and generate new feature combinations from an existing clinical dataset, and the non-parametric methods where virtual patients are produced by randomly selecting patients from a clinical dataset. Examples of parametric methods include the multivariate normal distribution (MVND) and its variant the multivariate log-normal distribution originally proposed by Tanenbaum et al. [3] towards the generation of virtual patients based on real clinical data. The MVND was also deployed in the work of Teutonico et al. [4] to create plausible virtual populations. A similar approach has been also introduced by RJ Allen [5], where the generated cohort data were able to match the observed data without the need for feature weighting. Silverman et al. [6] used multinomial logistic models to model sequence count data with complex covariance structure. Apart from the conventional statistical methods though, machine learning based methods have been also proposed. Bottcher et al. [7] developed a package named “deal” in R, which includes Bayesian networks for virtual population generation by taking into consideration the conditional probabilities among the features, supporting both discrete and continuous type of data. Robnik-Šikonja [8] utilized tree ensembles and artificial neural networks with radial basis functions (RBFs) as activation functions to detect hidden patterns among the features in the real data by either including or excluding a target feature yielding virtual data with decreased divergence with the real one.

None of the above virtual population generation studies have investigated the effectiveness of clinical data augmentation in terms of not only enhancing the size of the real patient data but also aggregating the virtually generated patient data with the real data to enhance the performance of disease classification and risk stratification models. In this work, we deploy five state-of-the-art virtual data generation methods to produce high-quality virtual patient data for 1000 patients with an increased level of similarity to the real patients across two clinical domains; the primary Sjogren’s Syndrome (pSS) and the hypertrophic cardiomyopathy (HCM). The number of virtually generated patients is relatively large for both clinical domains especially in pSS considering that it is a rare systemic autoimmune disease. The novelty of the proposed computational pipeline lies on the fact that it: (i) enhances the quality of the input clinical data through the precise detection and elimination of outliers and data inconsistencies using data curation workflows, (ii) augments the curated clinical data with high-quality

virtual data that enhance the population size of two rare clinical research databases through the development of high-performance virtual data generators, including both supervised and unsupervised tree ensembles, as well as, artificial neural networks (ANNs) with Gaussian kernels, which are extended to resolve overfitting effects during the generation stage, and (ii) builds supervised machine learning models on the aggregated real and virtual data for the robust classification of lymphoma patients with pSS and for the risk stratification of patients with HCM.

Our results highlight the favorable performance of the tree ensembles towards the generation of high-quality virtual data with goodness-of-fit (GOF) 0.021 and Kullback Leibler (KL)-divergence 0.029 in the pSS domain and 0.029, 0.027 in the HCM domain, respectively. The aggregation of the real and the virtual data from the tree ensembles revealed a notable increase in the classification accuracy, sensitivity, and specificity for both the lymphoma classification, where the XGBoost yielded an increase by 10.9% in accuracy, 10.7% in sensitivity, and 11.5% in specificity and the HCM risk stratification models, where the XGBoost yielded an increased by 16.1% in accuracy, 16.9% in sensitivity, and 13.7% in specificity. A similar increase is also observed in the case of the AdaBoost for HCM risk stratification (7.1% in accuracy, 5.7% in sensitivity, 10% in specificity, and 6.5% in AUC) and lymphoma classification (5.5% in the accuracy, 5.3% in sensitivity, 6.3% in specificity, and 10.1% in AUC), as well as, in the case of the Random Forests for HCM risk stratification (9.5% in accuracy, 8.9% sensitivity, 10.8% in specificity, and 11.2% in AUC) and lymphoma classification (9.4% in accuracy, 10.1% in sensitivity, 7.2% in specificity, and 12.2% in AUC). The outcomes of the proposed pipeline are promising since the existing lack of population size in both HCM and pSS obscure the development of robust disease classification and risk stratification models. To our knowledge, this is the first computational pipeline which aggregates high-quality virtual with real curated clinical data to address crucial clinical unmet needs in two rare clinical domains, including the development of robust lymphoma classification and HCM risk stratification models.

Section 2 presents the proposed computational pipeline along with the mathematical background of the methods for virtual population generation and machine learning. Section 3 presents the results of data augmentation in each clinical domain. The findings are discussed in Section 4.

## 2. Materials and methods

### 2.1. The proposed pipeline

The proposed pipeline for data augmentation is depicted in Fig. 1, which consists of three modules, namely the: (i) data quality control module for assessing the quality of the data, (ii) virtual population generation module for producing high-quality virtual data, and (iii) the “hybrid” machine learning module for the development of disease classification and risk stratification models on the aggregated real and virtual patient data. The outcomes of the proposed pipeline include curated clinical data, high-quality virtual data and enhanced disease classification and risk stratification models.

A data quality control pipeline presented in a previous study [9] was utilized to automatically resolve problematic fields within the input clinical data, including outliers, data inconsistencies, and missing values. The curated clinical data are introduced into the virtual population generation module to yield virtual distributions that “mimic” the real ones. Towards this direction, state-of-the-art machine-learning and statistical methods were developed to ensure the high-quality of the virtually generated data, including: (i) the supervised tree ensembles, where in each tree node, the generator for each feature is captured during the node splitting process based on its univariate empirical cumulative distribution function (ECDF), (ii) the unsupervised tree ensembles, where density forest ensembles are built in a top-down manner

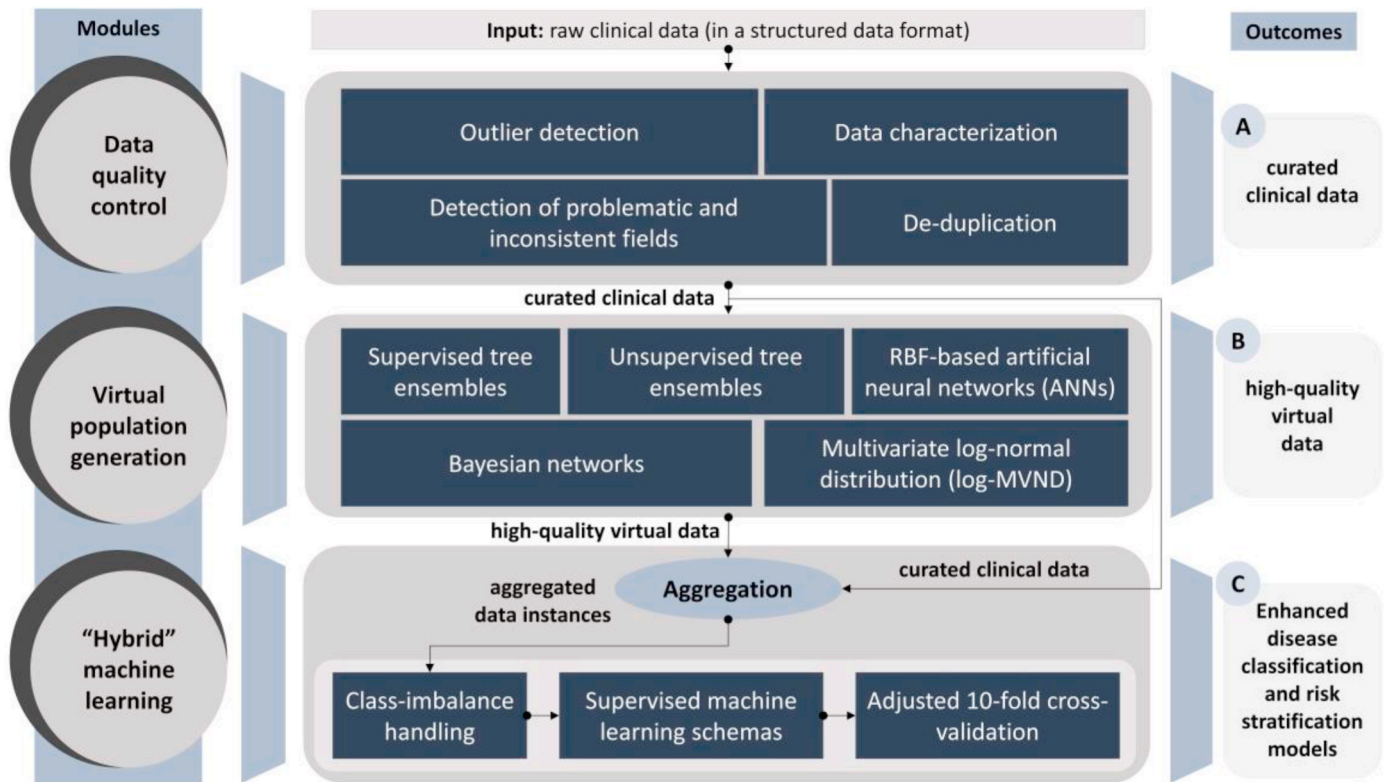


Fig. 1. An illustration of the proposed computational pipeline.

using the variance of the features as the criterion for the node splitting process, (iii) the artificial neural networks (ANNs), where radial basis functions (RBFs), such as, the Gaussian kernels are used as multivariate generators of virtual data instances, (iv) the Bayesian networks, where diverse network topologies are evaluated based on the causal relationships between the features (i.e., nodes in the network), and (v) the Log-MVND (multivariate log-normal distribution), where multivariate normal distributions are applied on the log transformed data. For each virtual population generation method, similarity scores, such as, the Kolmogorov-Smirnoff goodness of fit (GOF), the Kullback-Leibler (KL) divergence and the correlation coefficient are used to evaluate the level of agreement among the real and virtual distributions.

In the “hybrid” machine learning module, the virtual data from each generator are aggregated with the real data to assess whether the performance of the machine learning algorithms that are trained on the aggregated data is better than in the case where the algorithms are trained on the real data. Two case studies were conducted towards the development of robust lymphoma classification models in pSS and risk stratification models in HCM. Class imbalance handling was utilized to deal with the population imbalance among the control and target groups through the application of random downsampling with replacement on the control group. The XGBoost was deployed as a robust tree ensemble algorithm [12,13] which was trained on aggregated data instances along with the Adaptive Boosting (AdaBoost) [10] and the Random Forests [10] which were also deployed to evaluate the overall impact of data augmentation. An adjusted 10-fold cross validation procedure was utilized to train the algorithms on aggregated data instances and evaluate them on testing subsets of real patients.

## 2.2. Data quality control module

An improved version of a data quality control pipeline [9] was used to resolve incompatibilities and inconsistencies within the raw clinical data, including outliers, and duplicated features, aiming at improving the quality of the data in terms of completeness and conformity.

According to Fig. 1, the data quality control stage produces a data quality report, a diagnostic report, and a curated dataset. The data quality report includes feature-level meta-information regarding the data types, value ranges, and useful descriptive measures. In the diagnostic report, the data inconsistencies are marked using color coding [9]. The curated dataset is the original dataset, where data inconsistencies are resolved.

### 2.2.1. Data characterization

The features are annotated according to their data type as integer, float, and string or into unknown in the case where the data type is a mixture of multiple data types, as well as, into continuous or discrete. In case a feature has multiple missing values, the pipeline automatically marks it for removal.

### 2.2.2. Outlier detection

The z-score is a univariate approach for the detection of values having a large distance from their population mean [10]. Since the z-score might lead to misidentified outliers due to the non-robustness of the standard deviation, especially in small data samples, we use the modified z-score [10]:

$$z_{mod} = \frac{x - \tilde{x}}{MAD} = b \frac{x - \tilde{x}}{\text{median}(|x - \tilde{x}|)}, \quad (1)$$

where  $x$  is the feature vector,  $\tilde{x}$  is its mean value, MAD stands for the median absolute deviation,  $\tilde{x}$  is the median, and  $b$  is a correction factor that makes the MAD unbiased yielding robust results [10].

### 2.2.3. De-duplication

De-duplication involves the detection of potentially highly correlated features and/or lexically similar or identical features within the input data. In this work, we deploy the Spearman’s rank correlation [10] to identify features with increased correlation and the Levenshtein distance [10] to detect lexically similar features, as potential duplicates.

For two variables,  $c, d$ , the Spearman's rank correlation,  $s_{r,d}$ , is defined as:

$$s_{c,d} = \text{cov}(r_c, r_d) / (\sigma_{r_c} \sigma_{r_d}) \quad (2)$$

where  $r_c, r_d$  are the rank variables of  $c, d$ , respectively,  $\sigma_{r_c}, \sigma_{r_d}$  are the standard deviations of  $c, d$ , respectively and  $\text{cov}(r_c, r_d)$  is the covariance of  $r_c$  and  $r_d$ , respectively. The Levenshtein distance was used to measure the similarity between two strings,  $c$  and  $d$ , by taking into consideration the number of deletions, insertions, and substitutions that are needed to transform  $c$  into  $d$ :

$$\text{lev}_{c,d}(i,j) = \begin{cases} \max(i,j), & \min(i,j) = 0 \\ \min \begin{cases} \text{lev}_{c,d}(i-1,j) + 1 \\ \text{lev}_{c,d}(i,j-1) + 1 \\ \text{lev}_{c,d}(i-1,j-1) + 1_{(c_i \neq d_j)} \end{cases}, & \text{o.w.} \end{cases} \quad (3)$$

where 0 denotes that  $c, d$  are identical and values larger than 1 indicate the existence of differences.

### 2.3. Virtual population generation module

The virtual population generation stage includes five virtual data generation methods, both supervised and unsupervised, namely the: (i) the multivariate log-normal distribution (log-MVND), (ii) the supervised tree ensembles, (iii) the unsupervised tree ensembles, (iv) the RBF-based artificial neural networks (ANNs), and (v) the Bayesian networks. The implementation took place in Python 3.6 through the interconnection of R scripts from the packages “deal” [7] for the Bayesian networks and “semiArtificial” [8,28] for the tree ensembles and the ANNs. In this work, we generated 1000 virtual patients, which was a parameter setting of our generators.

#### 2.3.1. Multivariate log-normal distribution (log-MVND)

Given a univariate feature,  $X \in \mathbb{R}^{p \times n}$ , the multivariate normal distribution (MVND) can be defined as an extension of the normal distribution as in:

$$f(X) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} e^{-(X-\mu)\Sigma^{-1}(X-\mu)^T/2}, \quad (4)$$

where  $p$  is the dimension,  $\mu$  is the mean vector of  $X$ ,  $\Sigma$  is the covariance matrix of  $X$ , and  $\Sigma^{-1}$  is the pseudoinverse of  $\Sigma$ . A multi-dimensional normal distribution is constructed from the mean vector and the covariance matrix of the input data. To ease the assumption of normality within the data, the log-normal distribution is defined, where the logarithm of the exponential term in (4) fulfills the condition:

$$\ln(e^{f(x)}) \sim N(\mu, \Sigma). \quad (5)$$

#### 2.3.2. Supervised tree ensembles

A more advanced approach to virtual population generation is to train a tree ensemble [8,26–28] for a given set of training features and a target feature. During the training phase of the generator, we build an ensemble similar to random forests [26–28] with some additional data needed for data generation phase. In each interior tree node, we store the generator for the splitting feature based on its univariate empirical cumulative distribution function (ECDF). In each leaf node, we store ECDF-based generators for all variables not encountered on a path from the root to that leaf. To avoid overfitting effects which are introduced in the training process during the construction of the tree ensemble we introduce a new approach according to which one of the trees from the ensemble is randomly chosen when producing a new instance which is passed down the tree starting in the root node. The ensemble, as the generator, approximates the probability density function of the regions where features are assumed to be independent (the dependencies are likely to be resolved on the path from the root to the leaves). As the

ensemble contains a sufficient number of different trees, the probability density function of the original feature space is reasonable well approximated with the generated instances. During the training process, the Gini impurity index [8] is used to measure the probability of a variable,  $I$ , being classified in the wrong class:

$$I = 1 - \sum_{i=1}^n p_i^2, \quad (6)$$

where  $p_i$  is the probability of a sample falling in class  $i \in \{1, 2, \dots, k\}$ , and  $k$  is the number of classes.

#### 2.3.3. Unsupervised tree ensembles

The unsupervised tree ensemble generator is built in a similar way as the supervised tree ensemble, but instead of random forests ensemble, this generator builds a density forest ensemble [27]. Here, the ensemble members are density trees built with a similar top-down manner as decision trees but using the variance of the features as the criterion for selection of the splitting feature. To avoid overfitting effects which are introduced during the construction of the density forest ensemble, each density tree in the ensemble is randomly selected as the one with the smallest convergence rate. Any information regarding the target feature is not necessary. Other components are identical to the supervised tree ensemble both in the learning and generation phase.

RBF (Radial Basis Function) based artificial neural networks (ANNs).

Robnik-Šikonja [29] has proposed an approach for virtual population generation with artificial neural networks (ANNs), that uses radial base functions (RBFs) as activation functions. The RBF-based ANN's output is defined as in:

$$y(q) = \sum_{i=1}^N w_i \exp\left(-\beta \|q - q_i\|^2\right), \quad (7)$$

where  $y(q)$  is the output of the ANN,  $w_i$  is the weight of the  $i$ -th neuron,  $q_i$  is the center vector of the  $i$ -th neuron,  $\|q - q_i\|$  is the distance of each sample in  $q$  from the center vector  $q_i$  in the  $i$ -th neuron, and  $\beta$  is a standard Gaussian parameter. The RBF generator is created with a standard training algorithm which estimates the Gaussian parameters in the neurons. In the generation phase, the RBF generator uses Gaussian kernels as multivariate generators to deal with overfitting effects and produce new instances from each one in proportion to their presence in the training set.

#### 2.3.4. Bayesian networks

Bayesian networks are based on the idea that during learning, dependencies between variables are explicitly revealed and stored in a form of a directed acyclic graph (DAG). The nodes represent features and the edges connect nodes with a causal relationship. The weights on an edge determine the probability of ending up to a given value of a node given its predecessor. To generate new instances the structure is used to generate new feature values in a manner consistent with causal dependencies between the features. If the node is discrete, the probability distribution is uniform. If the node is continuous, one mean and variance is attached per configuration of the discrete parents.

Performance evaluation of the quality of the generated virtual patient data.

Many approaches on how to evaluate generators are shown in Ref. [8]. Below we present some of them.

#### 2.3.5. Goodness of fit (GOF)

The Kolmogorov-Smirnoff goodness of fit (GOF) test [14] is used to evaluate the similarity among the real and the virtual distributions. The GOF is defined as in:

$$g = \max(|f_r(x) - f_v(x)|), \quad (8)$$

where  $f_r(x)$  and  $f_v(x)$  are the empirical distribution functions of the



original and virtual data, respectively. A large gof value denotes distributions with large distance in at least a part of the distribution whereas a small one indicates small distance over the whole range of the distribution.

### 2.3.6. Pearson's correlation coefficient

For two variables, say  $x_r$  and  $x_v$ , coming from the real and virtual data, respectively, the correlation coefficient [15],  $r$ , is defined as in:

$$r = \frac{E[(x_r - \mu_{x_r})(y_r - \mu_{y_r})]}{\sigma_{x_r} \sigma_{y_r}}, \quad (9)$$

where  $\mu_{x_r}$ ,  $\mu_{y_r}$  are the mean values, and  $\sigma_{x_r}$ ,  $\sigma_{y_r}$  are the standard deviations of  $x_r$  and  $y_r$ , respectively, and  $E[\cdot]$  is the expectation operator.

### 2.3.7. Kullback-Leibler (KL) divergence

For two features,  $x_r$  and  $x_v$ , from the real and virtual data, respectively, with probability distributions,  $p_{x_r}$  and  $p_{x_v}$ , defined on the same probability space,  $K$ , the Kullback-Leibler (KL) divergence [16] quantifies the divergence between the two distributions, in an asymmetric manner, as in:

$$KL(p_{x_r} || p_{x_v}) = \sum_{k \in K} p_{x_r}(k_i) \log \left( \frac{p_{x_r}(k_i)}{p_{x_v}(k_i)} \right), \quad (10)$$

where KL values close to 0 denote that the two probability distributions  $p_{x_r}$  and  $p_{x_v}$  are almost identical in terms of highly reduced divergence or highly increased convergence.

## 2.4. Hybrid machine learning module

### 2.4.1. Class imbalance handling

Class imbalance handling was used to deal with the increased population imbalance among the control and target groups by randomly downsampling the control group with replacement into a 1:1 ratio. The process was repeated ten times, where on each round, the downsampled control group was matched according to age, gender, and disease duration with the target group.

### 2.4.2. Supervised machine learning schemas

The Extreme Gradient Boosting (XGBoost) algorithm was utilized as a state-of-the-art tree ensemble approach for the development of lymphoma classification models in pSS and risk stratification models in HCM. The algorithm is trained on aggregated instances from the real and virtual data and tested on the instances of the real patient data, where the accuracy, sensitivity, specificity, and area under the curve (AUC) are computed. In short, the XGBoost algorithm [12] uses classification and regression trees as base learners to sequentially combine multiple tree predictions through error minimization using gradient boosting thus yielding higher performance over the conventional decision trees and related bagging methods, such as, the random forests [12]. Given a set of  $M$ -features  $(u, v) = \{u_i, v_i\}$ ,  $i = 1, \dots, M$ , the XGBoost adds at a step  $t$ , a weak tree learner,  $f_t$ , that minimizes a regularized objective function,  $G(t)$ :

$$G(t) = \sum_{i=1}^M g\left(v_i, \tilde{v}_{i,t-1} + f_t(u_i)\right) + \gamma N + \frac{1}{2} \lambda \sum_{j=1}^J w_j^2, \quad (11)$$

where  $g(\cdot)$  is the loss function at step  $t$ ,  $\tilde{v}_{i,t-1}$  is the estimated value at step  $t - 1$ ,  $\gamma$  is a regularization term that avoids overfitting,  $w$  is the weight vector of the leaves,  $\lambda$  is a regularization scalar value, and  $N$  is the total number of leaves in each tree. To reduce the complexity, the first- and second-order gradients are computed according to Taylor's theorem yielding the objective function [17]:

$$G(t) = \sum_{j=1}^N \left[ FO_j w_j + \frac{1}{2} (SO_j + \lambda) w_j^2 \right] + \gamma N, \quad (12)$$

where  $FO_j = \sum_{i \in I_j} g_i$  and  $SO_j = \sum_{i \in I_j} h_i$ , are compact forms of the first- and second-order gradients,  $g_i = \partial l(y_i, \tilde{y}_{i,t-1}) / \partial \tilde{y}_{i,t-1}$  and  $h_i = \partial^2 l(y_i, \tilde{y}_{i,t-1}) / \partial \tilde{y}_{i,t-1}^2$  that are assigned to the  $j$ -th leaf.

Additional algorithms, including the Adaptive Boosting (AdaBoost) [10] and the Random Forests (RFs) [10] were deployed to evaluate the overall impact of data augmentation towards the development of robust lymphoma classification and HCM risk stratification models.

### 2.4.3. Adjusted 10-fold cross validation

To better understand the adjusted cross-validation process, let's denote the real dataset as  $T$ , and the virtual datasets that were generated by the methods from Section 2.3 as  $A$  for the unsupervised tree ensembles,  $B$  for the supervised tree ensembles,  $C$  for the supervised RBF-based neural networks,  $D$  for the Bayesian networks, and  $E$  for the multivariate log-normal distribution. A 10-fold cross validation process is applied to  $T$ , yielding a training subset  $T_{train}$  and a testing subset  $T_{test}$ , on each iteration. On each round, each virtual dataset is aggregated with  $T_{train}$ , yielding new training instances, say,  $A', B', C', D', E'$ , where  $A' = A \cup T_{train}$ ,  $B' = B \cup T_{train}$ ,  $C' = C \cup T_{train}$ ,  $D' = D \cup T_{train}$ ,  $E' = E \cup T_{train}$ . Each supervised machine learning algorithm from Section 2.4.2 is trained on the training instances  $A', B', C', D', E'$  and tested on the corresponding testing instance  $T_{test}$  where the accuracy, specificity, sensitivity, and area under the curve (AUC) scores are computed and averaged across the folds. In this way, the training instances of  $T$  are augmented with the virtual data.

## 3. Results

### 3.1. Data origins

We acquired two anonymized datasets. The first one consists of 449 patients who have been diagnosed with primary Sjögren's Syndrome (pSS) at the University of Athens (UoA) cohort. The number of lymphoma pSS patients was 70 with an average age 48.77 ( $\pm 12.54$ ) whereas the number of controls was 140 with an average age 52.47 ( $\pm 13.86$ ). There were 162 features, including demographics, medical conditions (e.g., dry eyes), and laboratory measures (e.g., C3), among others [19]. The second dataset includes 2454 records of patients who have been diagnosed with hypertrophic cardiomyopathy (HCM), at two time-points, from the Cardiomyopathies Unit at Careggi Hospital, Florence (UNIFI cohort) [20]. The number of high-risk patients was 300 with an average age 50.13 ( $\pm 17.67$ ) and the number of low-risk patients was 476 with an average age 43.95 ( $\pm 18.42$ ). There were 123 features, including demographics, laboratory measures (e.g., Left ventricular internal diameter end systole), and physical measures (e.g., systolic pressure), among others. All clinical data were shared according to the EU General Data Protection Regulation (GDPR) requirements [18].

### 3.2. Data preprocessing and quality control

According to Table 1, the pSS dataset from the UoA cohort consisted of 162 features (45% continuous, 36% discrete and 19% unknown). In total, 55% of the features had good quality in terms of reduced outliers and missing values, and 48% had bad quality and were removed from the pipeline. In addition, 19 features had inconsistent data formats and were also removed from the analysis. Outliers were detected in 16 features which were clinically examined and approved for correction. The final dataset included 449 patients with 65 features. To reduce the population imbalance among the lymphoma and non-lymphoma patients, which was approximately 6.4:1, the number of non-lymphoma

**Table 1**

A summary of the data quality report.

Metadata	UoA cohort	UNIFI cohort
Number of features	162	123
Number of records (instances)	449	2454
Number of discrete features	58	37
Number of continuous features	73	86
Number of unknown features	31	0
Number of features with outliers	16	0
Number of features with inconsistencies	19	0
Number of bad quality features	77	36
Number of fair quality features	57	42
Number of good quality features	26	45
Class imbalance ratio	1:6.4	1:1.58
Final number of patients after class imbalance handling	210	776
Final number of acceptable features	65	20

patients (i.e., the majority class) was set as twice the number of lymphoma patients. To do so the majority class was downsampled to preserve a 1:2 ratio among the lymphoma (presence = 1) and the non-lymphoma patients (lymphoma presence = 1, lymphoma absence = 0).

The HCM dataset from the UNIFI cohort (Table 1) consisted of 123 features (70% continuous, 30% discrete) and 2454 records. In total, 71% of the features had good quality, and 29% had bad quality status and were removed from the pipeline. No outliers were detected nor unknown data types or inconsistent fields. The class imbalance was acceptable in this case as the ratio of patients who have been diagnosed with high risk and low risk is approximately 1:1.58 (high-risk = 1; low-risk = 0). The number of missing values was 40% with more than 30 empty features which required prospective information. To avoid biases, the records were averaged across the two timepoints and the missing records were discarded yielding a final dataset of 776 records and 20 clinical features that were selected by the clinical experts.

### 3.3. Data augmentation

#### 3.3.1. Evaluation of the quality of the virtual data in pSS

The performance of the virtual data generators in the UOA cohort is presented in [Supplementary Table I](#) for the tree ensembles and the RBF-based ANNs, while in [Supplementary Table II](#) for the Bayesian networks and the log-MVND. The performance of the virtual generation methods was favorable. According to [Table 2](#), the average GOF was 0.021 for the unsupervised tree ensembles, 0.022 for the supervised tree ensembles, 0.068 for the RBF-based ANNs, 0.37 for the Bayesian networks and 0.133 for the Log-MVND. In addition, the average KL-divergence was 0.0289 for the unsupervised tree ensembles, 0.034 for the supervised tree ensembles, 0.033 for the RBF-based ANNs, 0.000005 for the Bayesian networks and 0.085 for the Log-MVND. The unsupervised tree ensembles generated virtual distributions with high similarity and convergence with the real data.

**Table 2**

Summary of the average performance evaluation measures for assessing the quality of the virtual data generated by each virtual population generation method for the pSS domain.

Virtual population generation method	Quality of the virtual data (goodness of fit, divergence, similarity)		
	GOF	KL-divergence	Correlation coefficient
Unsupervised tree ensembles	0.021	0.0289	0.1±0.22
Supervised tree ensembles	0.022	0.034	0.102±0.23
Supervised RBF-based ANNs	0.068	0.033	0.103±0.23
Bayesian networks	0.37	0.000005	0.06±0.07
Log-MVND	0.133	0.085	0.5±0.47

The absolute correlation difference between the real and virtual data by the unsupervised tree ensembles is depicted in [Fig. 2](#), where the average correlation difference was  $0.1 \pm 0.22$ . The white horizontal and vertical lines in the features “Renal disease” and “Kidney infiltrates” denote the existence of strong correlation differences. This occurs because only 4 patients had positive Renal disease while only 7 patients had positive kidney infiltrates among the 449 patients and thus the virtual distributions included only negative samples. The average correlation difference was  $0.102 \pm 0.23$  for the supervised tree ensembles,  $0.103 \pm 0.23$  for the RBF-based ANNs,  $0.5 \pm 0.47$  for the Log-MVND, and  $0.06 \pm 0.07$  for the Bayesian networks. The latter had the smallest correlation difference but lower GOF values than the unsupervised tree ensembles.

Random downsampling with replacement was applied on the real data to deal with the increased imbalance (1:6.4) among the lymphoma and non-lymphoma groups ([Table 1](#)), where the lymphoma over non-lymphoma ratio was set to 1:1. The process was repeated 10 times and the results were averaged across the random executions. On each execution, the downsampled group of non-lymphoma patients was matched according to age, gender, and disease duration with the group of lymphoma patients.

The application of the XGBoost on the real data yielded: accuracy = 0.724; sensitivity = 0.679; specificity = 0.814; AUC = 0.802. On the other hand, according to [Table 3](#), the average performance of the XGBoost on the aggregated real and virtual data from the unsupervised tree ensembles achieved the best classification performance, yielding accuracy 0.833, sensitivity 0.786, specificity 0.929, and AUC 0.924. The performance of the XGBoost using the augmented data from the supervised tree ensembles, and the supervised RBF-based ANNs come next. Finally, the performance of the XGBoost using the augmented data from the Log-MVND, and the Bayesian networks was lower than in the previous case (using the real data only).

In a similar manner, the performance of the lymphoma classification models from the AdaBoost and Random Forests using the augmented data from the tree ensembles was higher than in the case of the real data. The application of the AdaBoost on the real data yielded accuracy = 0.719, sensitivity = 0.675, specificity = 0.807, AUC = 0.749. On the other hand, according to [Table 3](#), the average performance of the AdaBoost on the aggregated real and virtual data from the unsupervised tree ensembles achieved better classification performance, yielding accuracy = 0.79, sensitivity = 0.732, specificity = 0.907, and AUC = 0.814.

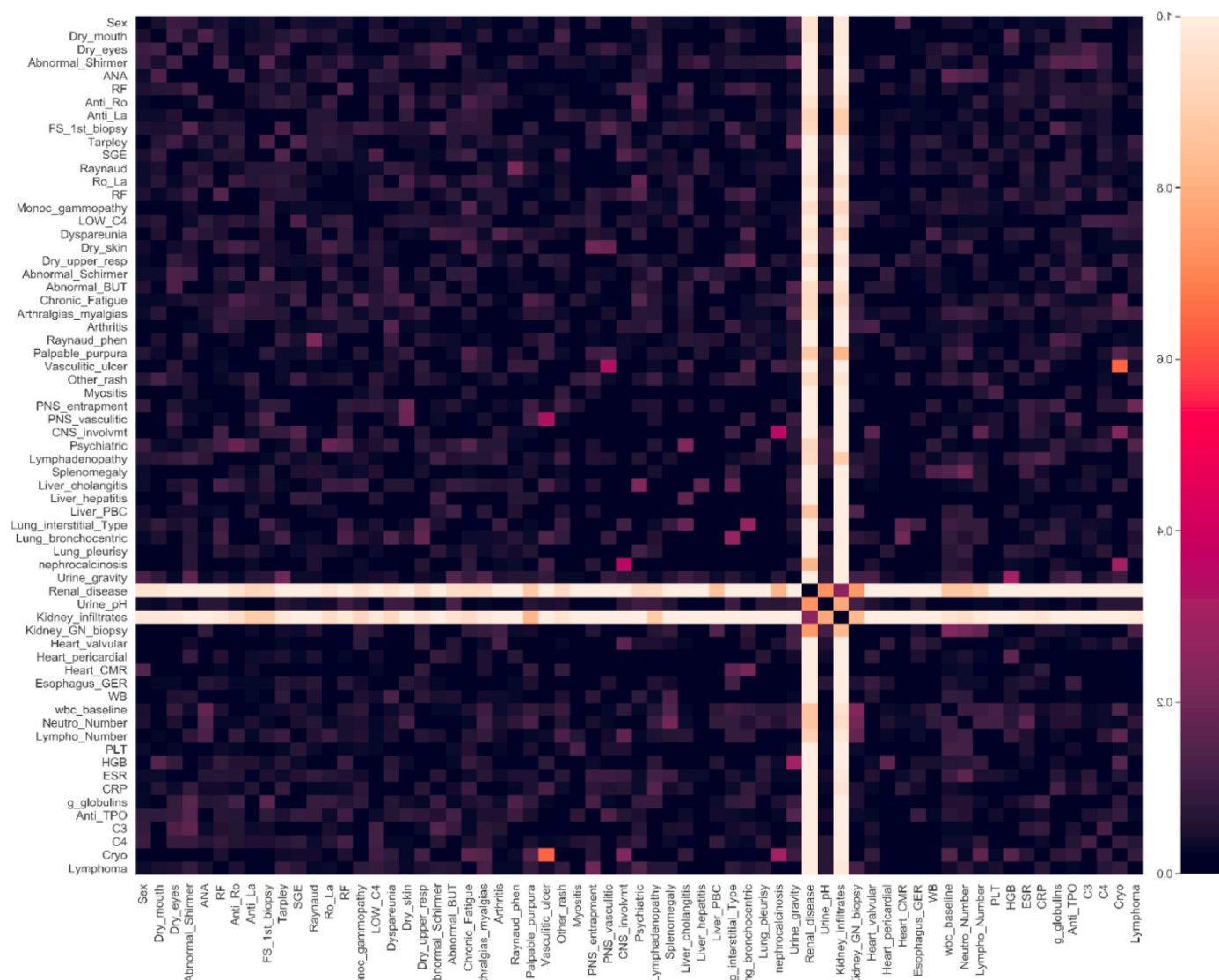
In the case of the Random Forests, the application on the real data yielded: accuracy = 0.729, sensitivity = 0.657, specificity = 0.871, AUC = 0.81. On the other hand, according to [Table 3](#), the average performance of the Random Forests on the aggregated real and virtual data from the unsupervised tree ensembles achieved better classification performance, yielding accuracy = 0.824, sensitivity = 0.746, specificity = 0.979, and AUC = 0.922.

The ROC curves are shown in [Fig. 3](#), highlighting the performance of the unsupervised tree ensembles (increase by 10.9% in the accuracy, 10.7% in sensitivity, 11.5% in specificity, and 12.2% in AUC) in the case of the XGBoost which suggests a notable performance enhancement. A similar increase is also observed in the case of the AdaBoost (7.1% in the accuracy, 5.7% in sensitivity, 10% in specificity, and 6.5% in AUC), as well as, in the case of the Random Forests (9.5% in the accuracy, 8.9% sensitivity, 10.8% in specificity, and 11.2% in AUC).

Evaluation of the quality of the virtual data in the case of hypertrophic cardiomyopathy.

The performance evaluation outcomes of the five virtual population generators in the HCM dataset are presented in [Supplementary Table III](#) for the unsupervised and supervised tree ensembles, and for the supervised RBF-based ANNs, while the Bayesian networks and the Log-MVND are shown in [Supplementary Table IV](#).

According to [Table 4](#), the average GOF was 0.029 for the unsupervised tree ensembles, 0.031 for the supervised tree ensembles, 0.23 for the RBF-based ANNs, 0.32 for the Bayesian networks and 0.198 for the



**Fig. 2.** The absolute difference between the real and virtual correlation matrices for the UoA dataset, in the case of the unsupervised tree ensembles generator. The features are ordered according to their appearance in the [Supplementary Table I](#). Values with dark and purple color denote low variations among the real and virtual data whereas values with orange/white color denote otherwise.

Log-MVND. The average KL-divergence was 0.027 for the unsupervised tree ensembles, 0.031 for the supervised tree ensembles, 0.02 for the RBF-based ANNs, 0.00047 for the Bayesian networks and 0.121 for the Log-MVND. The unsupervised tree ensembles generated virtual distributions with the highest similarity and reduced divergence with the real data.

The absolute correlation difference between the real and virtual data that were generated by the unsupervised tree ensembles is depicted in [Fig. 4](#), where the average difference was  $0.041 \pm 0.033$ . Regarding the rest of the algorithms, the average correlation difference was  $0.064 \pm 0.076$  for the supervised tree ensembles,  $0.078 \pm 0.085$  for the RBF-based ANNs,  $0.117 \pm 0.127$  for the Bayesian networks, and  $0.031 \pm 0.03$  for the Log-MVND. Although the Log-MVND schema achieved the smallest inter-correlation difference from the virtual population generators, it yielded significantly higher GOF and KL values than the unsupervised tree ensembles.

The dark color pattern in [Fig. 4](#) denotes the absence of significant correlation differences between the real and the virtual data which suggests that in this case the unsupervised tree ensembles schema was able to generate virtual distributions with increased similarity (i.e., with highly similar correlation patterns) with the real distributions.

In this case, class imbalance handling is not required since the ratio of the patients with low and high risk for HCM ([Table 1](#)) is adequate. The application of the XGBoost on the real data using a 10-fold cross validation process yielded accuracy = 0.597, sensitivity = 0.564, specificity = 0.708, and AUC = 0.628. According to [Table 5](#), the average performance of the XGBoost on the aggregated real and virtual data from the unsupervised tree ensembles achieved better classification performance, yielding accuracy = 0.758, sensitivity = 0.733, specificity = 0.845, and AUC = 0.829. The performance of the XGBoost on the augmented data from the supervised tree ensembles comes next along with the RBF-based ANNs, the Bayesian networks and the Log-MVND which achieved slightly better performance than before but with less than 0.6 sensitivity and thus are excluded from [Table 5](#).

The performance of the HCM risk stratification models from the AdaBoost and Random Forests using the augmented data from the tree ensembles was also higher than in the case of the real data. The application of the AdaBoost on the real data yielded accuracy = 0.61, sensitivity = 0.569, specificity = 0.748, and AUC = 0.611. According to [Table 5](#), the average performance of the AdaBoost on the aggregated data from the unsupervised tree ensembles achieved better classification performance, yielding accuracy = 0.665, sensitivity = 0.622, specificity



**Table 3**

A summary of the lymphoma classification results from the XGBoost, AdaBoost and Random Forests before and after data augmentation using the virtual data from each virtual population generation.

Virtual population generation method for data augmentation	Lymphoma classification performance			
	Accuracy	Sensitivity	specificity	AUC
<b>XGBoost</b>				
Before data augmentation	0.724	0.679	0.814	0.802
Unsupervised tree ensembles	0.833	0.786	0.929	0.924
Supervised tree ensembles	0.814	0.757	0.929	0.912
Supervised RBF-based ANNs	0.819	0.764	0.929	0.914
Bayesian networks	0.752	0.707	0.843	0.787
Log-MVND	0.8	0.754	0.893	0.824
<b>AdaBoost</b>				
Before data augmentation	0.719	0.675	0.807	0.749
Unsupervised tree ensembles	0.79	0.732	0.907	0.814
Supervised tree ensembles	0.79	0.725	0.921	0.82
Supervised RBF-based ANNs	0.824	0.764	0.943	0.87
Bayesian networks	0.69	0.593	0.886	0.76
Log-MVND	0.767	0.696	0.907	0.784
<b>Random Forests</b>				
Before data augmentation	0.729	0.657	0.871	0.81
Unsupervised tree ensembles	0.824	0.746	0.979	0.922
Supervised tree ensembles	0.767	0.661	0.979	0.877
Supervised RBF-based ANNs	0.757	0.636	1	0.901
Bayesian networks	0.762	0.661	0.964	0.839
Log-MVND	0.757	0.668	0.936	0.852

= 0.811, and AUC = 0.712. As for the Random Forests, their application on the real data yielded accuracy = 0.629, sensitivity = 0.563, specificity = 0.853, AUC = 0.641, whereas the average performance on the aggregated real and virtual data from the unsupervised tree ensembles achieved better classification performance, yielding accuracy = 0.723, sensitivity = 0.664, specificity = 0.925, and AUC = 0.763.

The ROC curves are summarized in Fig. 5, highlighting the classification performance of the unsupervised tree ensembles which yielded an

increase by 16.1% in the accuracy, 16.9% in sensitivity, 13.7% in specificity, and 20.1% in AUC compared with the XGBoost trained on the real data. A similar increase is also observed in the case of the AdaBoost (5.5% in accuracy, 5.3% in sensitivity, 6.3% in specificity, and 10.1% in AUC), as well as, in the Random Forests (9.4% in accuracy, 10.1% in sensitivity, 7.2% in specificity, and 12.2% in AUC). Although the classification performance is significantly smaller than in pSS, due to the nature of the HCM, data augmentation was able to enhance the performance of the HCM risk stratification models.

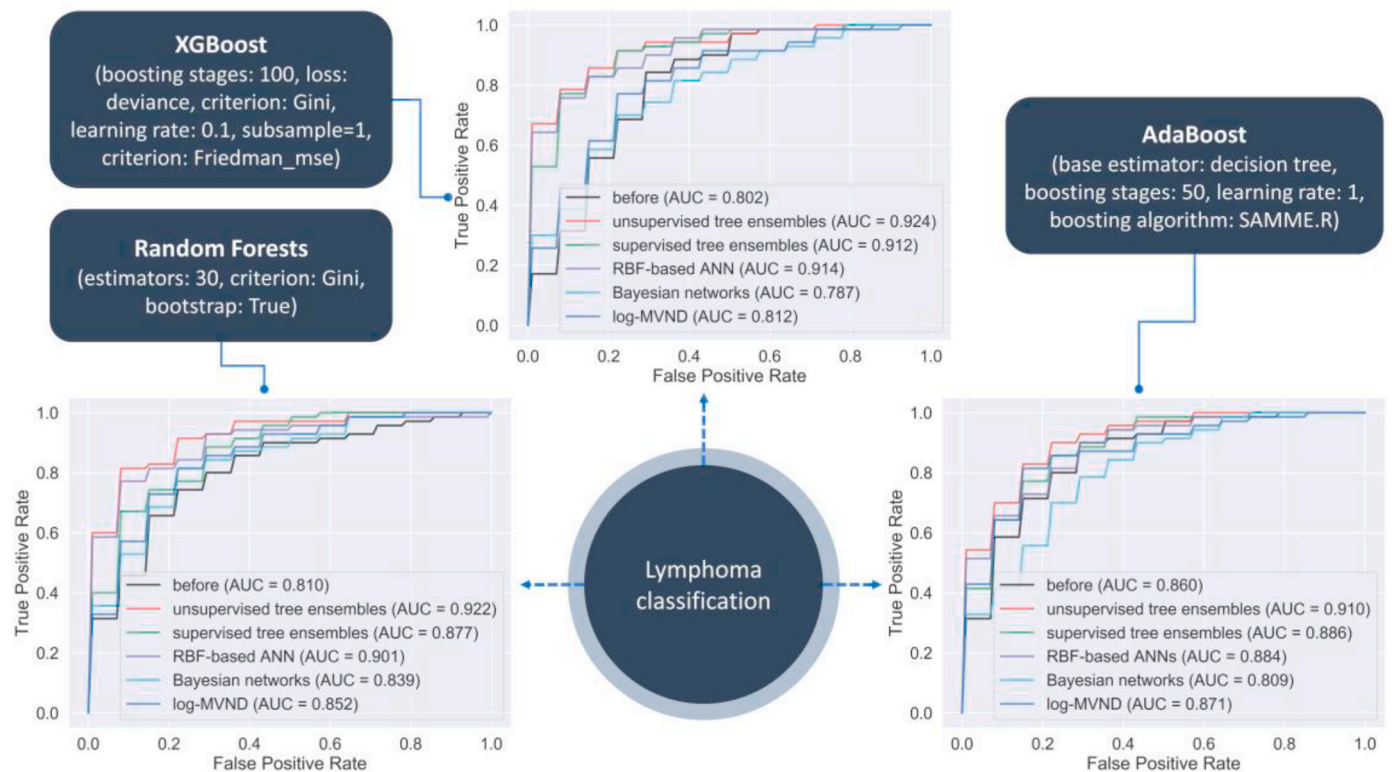
#### 4. Discussion

In this work we examined the effectiveness of data augmentation in terms of enhancing the real clinical research databases with high-quality virtual data to enhance the performance of the disease classification and risk stratification models in two different clinical domains, namely the primary Sjögren's Syndrome and the hypertrophic cardiomyopathy. To do so, a computational pipeline was developed, where high-quality

**Table 4**

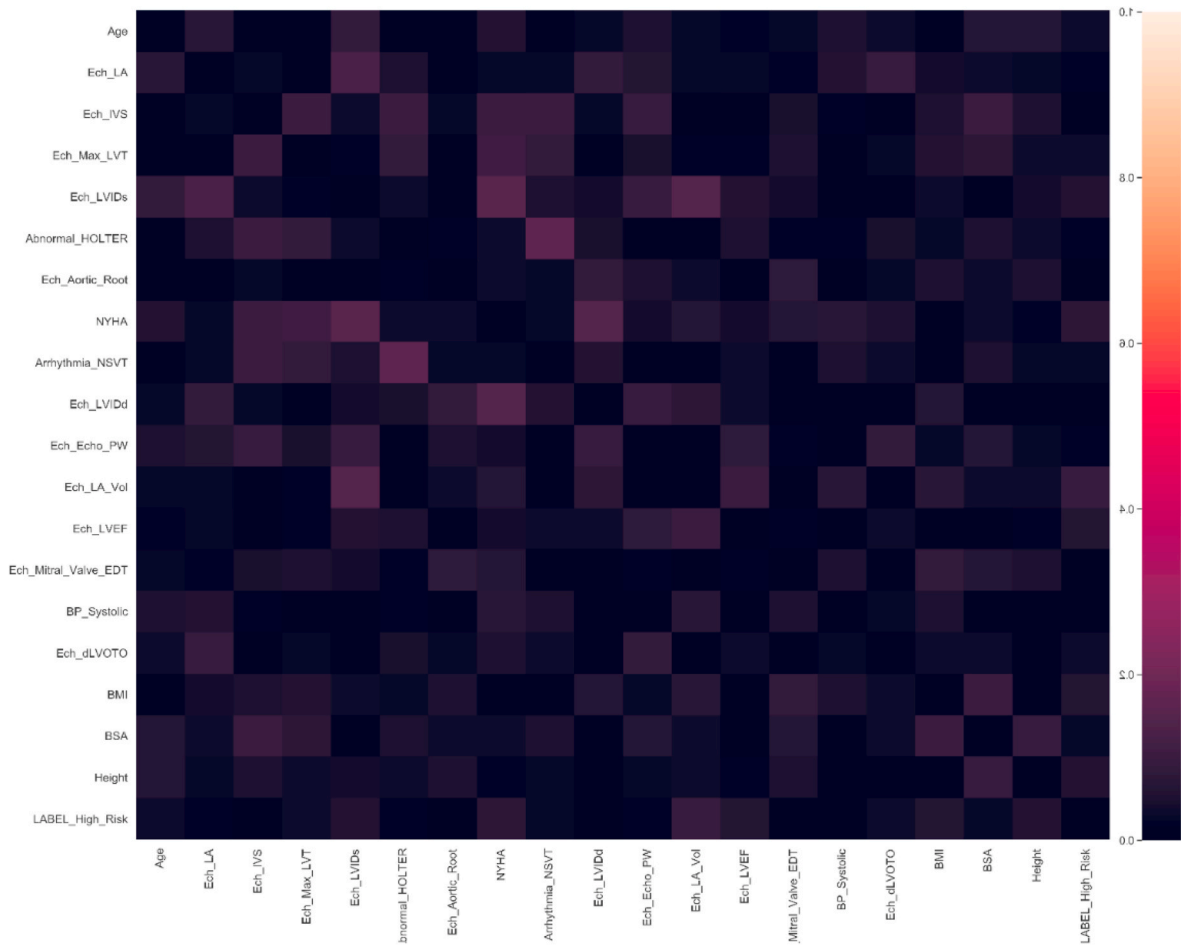
Summary of the average performance evaluation measures for assessing the quality of the virtual data generated by each virtual population generation method for the HCM domain.

Virtual population generation method	Quality of the virtual data (goodness of fit, divergence, similarity)		
	GOF	KL-divergence	Correlation coefficient
Unsupervised tree ensembles	0.029	0.027	0.041±0.033
Supervised tree ensembles	0.031	0.031	0.064±0.076
Supervised RBF-based ANNs	0.23	0.02	0.078±0.085
Bayesian networks	0.32	0.0047	0.117±0.127
Log-MVND	0.198	0.121	0.031±0.03



**Fig. 3.** ROC curves depicting the classification performance of the XGBoost, the AdaBoost and the Random Forests for lymphoma classification with and without data augmentation.





**Fig. 4.** The absolute difference between the real and virtual correlation matrices for the HCM dataset, in the case of the unsupervised tree ensembles generator. The features are ordered based on their appearance in [Supplementary Table III](#). Risk stratification for hypertrophic cardiomyopathy using data augmentation.

**Table 5**

A summary of the HCM risk stratification results from the XGBoost, AdaBoost and Random Forests before and after data augmentation using the virtual data from each virtual population generator.

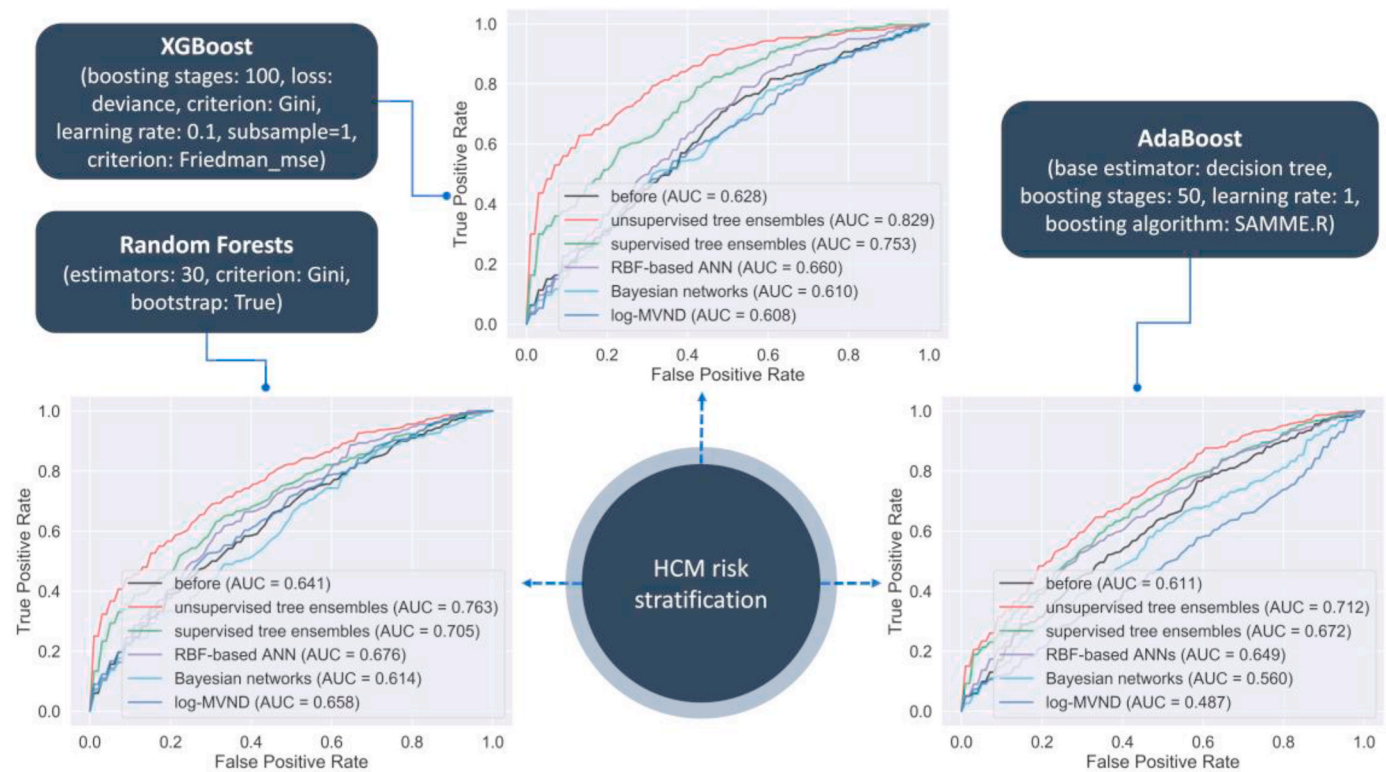
Virtual population generation method for data augmentation	HCM risk stratification performance			
	Accuracy	sensitivity	specificity	AUC
<b>XGBoost</b>				
Before data augmentation	0.597	0.564	0.708	0.628
Unsupervised tree ensembles	0.758	0.733	0.845	0.829
Supervised tree ensembles	0.705	0.672	0.817	0.753
<b>AdaBoost</b>				
Before data augmentation	0.61	0.569	0.748	0.611
Unsupervised tree ensembles	0.665	0.622	0.811	0.712
Supervised tree ensembles	0.653	0.606	0.816	0.672
<b>Random Forests</b>				
Before data augmentation	0.629	0.563	0.853	0.641
Unsupervised tree ensembles	0.723	0.664	0.925	0.763
Supervised tree ensembles	0.686	0.621	0.908	0.705

virtual data are aggregated with the real data to yield robust lymphoma classification models and HCM risk stratification models, where the performance of each model was evaluated on testing instances of the real data to avoid any biases. The proposed pipeline was able to generate virtual distributions with increased similarity, correlation, and reduced divergence with the real distributions. The aggregation of the real with the virtual patient data in both clinical domains yielded a notable increase in the classification accuracy, sensitivity, specificity, and area under the curve scores of the supervised machine learning models which

were trained on the augmented clinical data compared to those trained on real data instances.

Our results reveal the favorable performance of the unsupervised tree ensembles for virtual population generation which outperformed the rest of the virtual population generation methods having the smallest goodness of fit and Kullback-Leibler divergence values in both experimental case studies (see [Table 2](#) regarding the pSS domain and [Table 5](#) regarding the HCM domain). The histograms of the virtual data that were generated by the unsupervised tree ensembles can be found in [Supplementary Figs. 1 and 2](#) for the HCM dataset and in [Supplementary Figs. 3 and 4](#) for the pSS dataset. In both cases, the histograms reflect a highly qualitative similarity between the real and the virtual distributions. The supervised tree ensembles had the second-best performance ([Tables 2 and 5](#)). The results from the supervised RBF-based artificial neural networks (ANNs) are close to the two previous methods, with the Bayesian networks and the log-MVND trailing behind. The dominance of the tree ensembles as a method for generating virtual data with increased level of agreement with the real data is in line with a recent study [21] which focuses on the generation of virtual data for in-silico cardiomyopathies drug development as an extension of [22].

Our results also highlight the positive impact of augmenting the real with the virtual patient data which were generated by the “unsupervised” tree ensembles through data augmentation towards the development of robust disease classification and risk stratification models. The XGBoost algorithm was selected as a state-of-the-art tree ensemble approach the value of which was demonstrated in previous studies [23, 24] for lymphoma classification in pSS. The performance of the lymphoma classification model in the pSS domain showed an increase by



**Fig. 5.** ROC curves depicting the classification performance of the XGBoost, the AdaBoost and the Random Forests for HCM risk stratification with and without data augmentation.

10.9% in the classification accuracy, 10.7% in sensitivity, 11.5% in specificity, and 12.2% in area under a curve for lymphoma classification (Fig. 3, Table 3) against the one trained only on the real data. A similar increase is also observed in the case of the AdaBoost (7.1% in accuracy, 5.7% in sensitivity, 10% in specificity, and 6.5% in AUC), as well as, in the case of the Random Forests (9.5% in the accuracy, 8.9% sensitivity, 10.8% in specificity, and 11.2% in AUC).

Moreover, the performance of the HCM risk stratification model showed an increase by accuracy, 16.9% in sensitivity, 13.7% in specificity, and 20.1% in area under the curve against the one trained on the real HCM data (Fig. 5, Table 5). A similar increase is also observed in the case of the AdaBoost (5.5% in accuracy, 5.3% in sensitivity, 6.3% in specificity, and 10.1% in AUC), as well as, in the case of the Random Forests (9.4% in accuracy, 10.1% in sensitivity, 7.2% in specificity, and 12.2% in AUC). In addition, the aggregation of the virtual data from the supervised tree ensembles with the real patient data yielded enhanced classification models at a similar extent (see Fig. 3, Table 3 for lymphoma classification and Fig. 5, Table 5 for HCM risk stratification). Finally, the aggregation of the virtual data from the supervised RBF-based ANNs, the Bayesian networks and the Log-MVND with the real one yielded supervised machine learning models with partially enhanced performance while maintaining the increased performance than in the case of training on the real data only.

## 5. Existing work/Contribution to the state of the art

Our work contributes positively towards the generation of high-quality virtual data that mimic the real data with an increased level of similarity and can be applied for data augmentation purposes to increase the population size of clinical studies, as well as, increase the robustness of the existing disease classification and risk stratification models. The current study builds on principles from existing studies (Table 6) to develop a beyond the state-of-the art computational pipeline for clinical data augmentation. We extend the conventional statistical approaches,

**Table 6**  
Contribution to the state-of-the art.

Study	Proposed method for virtual population generation towards data augmentation/Rationale
Teutonico et al. [4]	Multivariate and discrete re-sampling techniques to account for covariate effects within the target population during the generation of virtual data.
Allen et al. [5]	A technique for efficiently generating virtual patients that best fit the observed data using multivariate log-normal distribution (log-MVND).
Tannenbaum et al. [3]	Continuous and categorical covariate distribution modeling using multivariate statistical functions.
Silverman et al. [6]	Application of logit logistic normal multivariate methods for population.
Krauss et al. [25]	Application of Bayesian methods for virtual population physiologically based Pharmacokinetic (PPBK) and pharmacokinetic modeling.
Robnik-Šikonja [28,29]	Definition of the tree ensembles and the RBF-based ANNs for virtual population generation yielding virtual data that mimic the UCI data.
Current study	A computational pipeline with properly designed machine learning methods for data augmentation which significantly enhances the performance of disease classification and risk stratification models across two different clinical domains through semi-supervised learning.

such as, the MVND and the Log-MVND [3–5], as well as, multivariate functions, such as, Bayesian methods, discrete re-sampling techniques [5,6,25], through machine learning based generators, such as, the tree ensembles, the RBF-based ANNs and the Bayesian networks to produce high-quality virtual patient data with increased similarity and decreased divergence with the real patient data.

All in all, our results validate the scientific and technical impact of data augmentation in both clinical domains yielding a significant increase in the classification accuracy, sensitivity, and specificity for both the lymphoma classification and the HCM risk stratification models. To

our knowledge, this is the first study that builds a computational pipeline which uses the high-quality semi-artificial patient data which are generated by machine learning-based approaches to enhance the performance of lymphoma classification models in pSS and risk stratification models in HCM.

## 6. Conclusions

In this work, we deploy high-quality virtually generated patient data to enhance the performance of the conventional supervised machine learning models for lymphoma classification and HCM risk stratification in two rare clinical domains. The proposed computational pipeline can be deployed for the augmentation of clinical data although medical imaging information can also be used as input. To our knowledge, this is the first computational pipeline which aggregates high-quality virtual data with real data to deal with clinical unmet needs in two rare clinical domains, including the development of robust lymphoma classification and HCM risk stratification models. The data quality control module enhances the quality of the raw clinical data through the removal of outliers and duplicated fields. The virtual population generation module provides straightforward virtual data generators, where the tree ensemble generators have been extended to avoid overfitting effects during the generation stage yielding virtual data with increased quality in terms of increased convergence with the real data. A similar strategy was developed for the ANNs using Gaussian kernels as activation functions to deal with overfitting during the training stage. The “hybrid” machine learning module utilizes supervised machine learning algorithms on the aggregated real and high-quality virtual data to enhance the performance of the lymphoma classification and HCM risk stratification models.

Although the application of the proposed pipeline has a strong potential towards the improvement of the existing disease classification and risk stratification models in other clinical domains, emphasis must be given on its concise application in each domain. The data quality control module enhances the quality of the input data by removing data inconsistencies and incompatibilities, but it should be utilized prior to the application of the virtual population generation module otherwise the quality of the virtual data will be poor in terms of reduced conformity and relevance with the real data. In addition, although the virtual population generators and specifically the tree ensembles and the ANNs have been adjusted to resolve overfitting effects during the training stage, emphasis should be given on the precise definition of the data types of the input features to avoid the generation of virtual data with heterogeneous data structure. The quality of the virtual data should be evaluated in terms of increased similarity and reduced divergence with the real data, where only the virtual data with the highest quality should be augmented with the real data. Finally, the statistical power of the augmented clinical data must be sufficient for the application of the hybrid machine learning module to yield robust disease classification and risk stratification models.

## 7. Future work

As a feature work, we plan to apply the proposed computational pipeline in other clinical domains to enhance the population size of clinical research databases with reduced statistical power. In addition, we plan to expand the hybrid machine learning module with deep learning algorithms to support the extraction of biomarkers from time-series gene expression data, as well as, enhance the applicability of the data quality control module to support the curation of complex genetic data structures.

## Author contributions

VCP developed the virtual population generation workflows and conducted the overall lymphoma classification and HCM risk

stratification analysis with and without the augmented data. VCP, GIG, and NS prepared the Introduction. GIG, GG, and NS prepared the figures and contributed to the writing of the Materials and Methods and Results. IO and FB contributed to the writing of the data description for the HCM domain, as well as, to the evaluation of the quality of the virtually generated data and the performance of the AI models for HCM risk stratification, as well as, to the interpretation of the findings in the Discussion. AG and AGG contributed to the writing of the data description for the pSS domain, to the evaluation of the quality of the virtually generated data and the performance of the AI models for lymphoma classification, as well as, to the interpretation of the findings in the Discussion. TS, ZB, MP, MRS, DGJ, and DIF had major contribution to the structuring of the Material and Methods and Results, as well as, to the detailed review of the manuscript.

## Acknowledgement

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 777204. This paper reflects only the author’s view and the Commission is not responsible for any use that may be made of the information it contains.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.combiomed.2021.104520>.

## References

- [1] M. Viceconti, A. Henney, E. Morley-Fletcher, In silico clinical trials: how computer simulation will transform the biomedical industry, *International Journal of Clinical Trials* 3 (2) (2016) 37–46.
- [2] E.D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, Q.V. Le, Autoaugment: learning augmentation strategies from data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 113–123.
- [3] S.J. Tannenbaum, N.H.G. Holford, H. Lee, C.C. Peck, D.R. Mould, Simulation of correlated continuous and categorical variables using a single multivariate distribution, *J. Pharmacokinet. Pharmacodyn.* 33 (6) (2006) 773–794.
- [4] D. Teutonico, F. Musuamba, H.J. Maas, A. Facius, S. Yang, M. Danhof, O. Della Pasqua, Generating virtual patients by multivariate and discrete Re-sampling techniques, *Pharmaceut. Res.* 32 (10) (2015) 3228–3237.
- [5] R. Allen, T. Rieger, C. Musante, Efficient generation and selection of virtual populations in quantitative systems pharmacology models: generation and selection of virtual populations, *CPT Pharmacometrics Syst. Pharmacol.* 5 (3) (2016) 140–146.
- [6] J. D. Silverman, K. Roche, Z.C. Holmes, L.A. David, S. Mukherjee, Bayesian Multinomial Logistic Normal Models through Marginally Latent Matrix-T Processes, 2019 arXiv preprint arXiv:1903.11695.
- [7] S.G. Böttcher, C. Dethlefsen, deal: a package for learning Bayesian networks, Documentation available in, <https://cran.r-project.org/web/packages/deal/deal.pdf>. (Accessed October 2018).
- [8] M. Robnik-Sikonja, Dataset comparison workflows, *International Journal of Data Science* 3 (2) (2018) 126.
- [9] V.C. Pezoulas, K.D. Kourou, F. Kalatzis, T.P. Exarchos, A. Venetsanopoulou, E. Zampeli, S. Gandolfo, F. Skopouli, S. De Vita, A.G. Tzioufas, D.I. Fotiadis, Medical data quality assessment: on the development of an automated framework for medical data curation, *Comput. Biol. Med.* 107 (2019) 270–283.
- [10] V.C. Pezoulas, T.P. Exarchos, D.I. Fotiadis, “Medical Data Sharing, Harmonization and Analytics – Chapter 3 Medical Data Sharing,” Academic Press, Paperback, 2020. ISBN: 9780128165072, eBook ISBN: 9780128165591.
- [11] V.C. Pezoulas, K.D. Kourou, F. Kalatzis, T.P. Exarchos, A.I. Venetsanopoulou, E. Zampeli, S. Gandolfo, F.N. Skopouli, S. De Vita, A.G. Tzioufas, D.I. Fotiadis, “Enhancing medical data quality through data curation: a case study in primary Sjögren’s syndrome, *Clin. Exp. Rheumatol.* 37 (3) (2019) 90–96.
- [12] V.C. Pezoulas, F. Kalatzis, T.P. Exarchos, A. Goules, S. Gandolfo, E. Zampeli, F. Skopouli, S. De Vita, A.G. Tzioufas, D.I. Fotiadis, “Dealing with Open Issues and Unmet Needs in Healthcare through Ontology Matching and Federated Learning,” Accepted for Presentation in the 2020 8th European Medical and Biological Engineering Conference, (EMBECE), 2020.
- [13] T. Chen, T. He, M. Benesty, V. Khotilovich, Y. Tang, Xgboost: extreme gradient boosting, R package version 0 4–2 (2015) 1–4.
- [14] R.B. D’Agostino, M.A. Stephens (Eds.), *Goodness-of-fit Techniques*, Marcel Dekker, Inc., New York, NY, USA, 1986.
- [15] P. Schober, C. Boer, L.A. Schwarte, Correlation coefficients: appropriate use and interpretation, *Anesth. Analg.* 126 (5) (2018) 1763–1768.

- [16] Y. Bu, S. Zou, Y. Liang, V. Veeravalli, Estimation of KL divergence: optimal minimax rate, *IEEE Trans. Inf. Theor.* 64 (4) (2018) 2648–2674.
- [17] J.H. Friedman, Greedy function approximation: a gradient boosting machine, *Ann. Stat.* 29 (2001) 1189–1232.
- [18] Regulation (Eu), 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), *Off J Eur Union* 119 (2016) 1–88.
- [19] S. Fragkioudaki, C.P. Mavragani, H.M. Moutsopoulos, Predicting the risk for lymphoma development in Sjogren syndrome: an easy tool for clinical use, *Medicine* 95 (25) (2016), e3766.
- [20] F. Mazzarotto, F. Girolami, B. Boschi, F. Barlocco, A. Tomberli, K. Baldini, I. Olivotto, Defining the diagnostic effectiveness of genes for inclusion in panels: the experience of two decades of genetic testing for hypertrophic cardiomyopathy at a single center, *Genet. Med.* 21 (2) (2019) 284–292.
- [21] V.C. Pezoulas, G.I. Grigoriadis, N.S. Tachos, I. Olivotto, D.I. Fotiadis, Generation of virtual patient data for in silico cardiomyopathies drug development using tree ensembles: a comparative study. Accepted for Oral Presentation at the 2020 IEEE 43rd International Conference on Engineering in Medicine and Biology, 2020.
- [22] V.C. Pezoulas, N. Tachos, D.I. Fotiadis, Generation of virtual patients for in silico cardiomyopathies drug development. In 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering, (BIBE), 2019, pp. 671–674, <https://doi.org/10.1109/BIBE.2019.00126>.
- [23] V.C. Pezoulas, K.D. Kourou, T. Kalatzis, T.P. Exarchos, E. Zampeli, S. Gandolfo, A. Goules, C. Baldini, F. Skopouli, S. De Vita, A.G. Tzioufas, D.I. Fotiadis, Overcoming the barriers that obscure the interlinking and analysis of clinical data through harmonization and incremental learning, *IEEE Open Journal of Engineering in Medicine and Biology* 1 (2020) 83–90.
- [24] V.C. Pezoulas, T.P. Exarchos, A. G. Tzioufas, S. De Vita, D.I. Fotiadis, “Predicting lymphoma outcomes and risk factors in patients with primary Sjögren’s Syndrome using gradient boosting tree ensembles,”. Proceedings of the 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, (EMBC), 2019, pp. 2165–2168.
- [25] M. Krauss, A. Schuppert, Assessing interindividual variability by Bayesian-PBPK modeling, *Drug Discov. Today Dis. Model.* 22 (2016) 15–19.
- [26] Parikshit Ram, Alexander G. Gray, Density estimation trees. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD, 2011, pp. 627–635. ACM.
- [27] Leo Breiman, Random forests, *Machine Learning Journal* 45 (5–32) (2001).
- [28] Marko Robnik-Šikonja, *semiArtificial: Generator of Semi-artificial Data*, 2014. URL, <http://cran.r-project.org/package=semiArtificial>. R package version 1.2.0.
- [29] Marko Robnik-Šikonja, Data generators for learning systems based on RBF networks, *IEEE Transactions on Neural Networks and Learning Systems* 27 (5) (May 2016) 926–938.