2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Variational Gaussian Mixture Models with robust Dirichlet concentration priors for virtual population generation in hypertrophic cardiomyopathy: a comparison study

Vasileios C. Pezoulas, *Student Member, IEEE,* Grigorios I. Grigoriadis, Nikolaos S. Tachos, *Member IEEE*, Fausto Barlocco, Iacopo Olivotto, and Dimitrios I. Fotiadis, *Fellow, IEEE*

***Abstract*— Nowadays, there is a growing need for the development of computationally efficient virtual population generators for large-scale *in-silico* clinical trials. In this work, we utilize the Gaussian Mixture Models (GMM) with variational Bayesian inference (BGMM) using robust estimations of Dirichlet concentration priors for the generation of virtual populations. The estimations were based on an exponential transformation of the number of Gaussian components. The proposed method was compared against state-of-the-art virtual data generators, such as, the Bayesian networks, the supervised tree ensembles (STE), the unsupervised tree ensembles (UTE), and the artificial neural networks (ANN) towards the generation of 20000 virtual patients with hypertrophic cardiomyopathy (HCM). Our results suggest that the proposed BGMM can yield virtual distributions with small inter- and intra-correlation difference (0.013 and 0.012), in lower execution time (4.321 sec) than STE which achieved the second-best performance.**

Keywords: Virtual population, Gaussian Mixture Models, variational inference, Hypertrophic cardiomyopathy

I. Introduction

In the recent years, there is an emerging need for virtual population generation in *in-silico* clinical trials (ISCTs), where the financial burden of expensive drug testing and development is large [1-4]. Virtual population generation is a computational approach which can provide insight into the pathogenic mechanisms of different diseases, such as, the cardiovascular diseases (CVDs) through the augmentation of real patient data with high-quality virtual patient data that "mimic" the real ones. So far, virtual population generation has multiple applications in ISCTs specifically in drug testing and development [1, 3], as well as, in pharmacokinetics [2, 4].

Probabilistic approaches are the most common methods for virtual population generation [5, 6], where the synthetic samples are randomly drawn from the real distributions. A widely used probabilistic method is the multivariate normal distribution (MVND) which has been widely adopted for the generation of virtual patients in *in-silico* clinical trials, such as, in hypertrophic cardiomyopathy (HCM) [6]. MVND utilizes multi-dimensional normal distributions given the mean vector and the covariance matrix of the real data to generate synthetic data. However, a fundamental assumption in MVND is that the real data follow a normal distribution. Bayesian networks [7] have been also used for the generation of virtual data based on conditional probabilities across different network topologies, where the nodes represent the features. Another family of virtual population generation involves the application of machine learning algorithms, such as, the supervised tree ensembles (STE), the unsupervised tree ensembles (UTE) [8-10], and the artificial neural networks (ANNs) with radial basis functions [8-10] which are trained on the real data and then transformed into data generators.

The emerging need for the development of computationally efficient virtual data generators yielding virtual data with reduced inter- and intra- correlation with the real data remains a technical challenge. The state-of-the-art virtual generators yield high-quality virtual data with reduced goodness of fit (gof) values, like the UTE [8]. The gof, however, assumes that the distributions belong to a particular set of distributions [11] which introduces biases in the outcomes. In addition, the STE, and the ANN [9, 10] require a target feature which affects the associations of the features in the virtual data. Furthermore, in the case of Bayesian networks, the number of all possible permutations of the edges within the network is infinite [7]. Moreover, the majority of these methods are computationally demanding due to the increased training time.

Towards this direction, Gaussian Mixture Models (GMMs) with variational Bayesian inference (BGMM) were developed to generate large-scale virtual populations. The proposed method utilizes Dirichlet process mixtures as the BGMM's prior structure, where the concentration of each component on the weight distribution is an exponential function of the number of components. Our approach was compared against state-of-the-art virtual data generators, including the Bayesian networks, the STE, the UTE, and the ANN for the generation of 20000 virtual patients for *in-silico* clinical trials in HCM yielding the lowest inter- and intra-correlation differences (0.013 and 0.012), in lower execution time (4.321) than the STE (46.537 sec) which had the second-best performance.

*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.

V.C. Pezoulas, G.I. Grigoriadis and N.S. Tachos are with the Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, University of Ioannina, GR 45110 Ioannina, Greece (e-mails: bpezoulas@gmail.com, greg8grigoriadis@gmail.com, ntachos@gmail.com).

D.I. Fotiadis is with the Department of Biomedical Research, Institute of Molecular Biology and Biotechnology, FORTH, Ioannina, Greece and the Dept. of Materials Science and Engineering, Unit of Medical Technology and Intelligent Information Systems, University of Ioannina, GR 45110, Ioannina, Greece (e-mail: fotiadis@uoi.gr).

F. Barlocco and I. Olivotto are with the Department of Experimental and Clinical Medicine, University of Florence and Cardiomyopathies Unit, Azienda Ospedaliera Careggi, Florence, Italy (e-mails: fausto.barlocco@unifi.it, iacopo.olivotto@gmail.com).

II. Materials And Methods

A. Data sharing and data quality control

Anonymized data were obtained from 776 patients under the SILICOFCM project [12]. The dataset included 20 features related to demographic and echocardiographic measurements. A data curation pipeline presented in a previous study [13] was applied on the clinical data to remove outliers, duplicated fields, and inconsistent data types using both univariate and multivariate methods [13].

B. Methods for virtual population generation

1) Bayesian networks

Bayesian networks [14] are based on the definition of a directed acyclic graph (DAG), say $\boldsymbol{D} = (\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V}$ is a set of nodes and $\boldsymbol{E}$ is a set of directed edges between the nodes in $\boldsymbol{V}$. Each node $v \in \boldsymbol{V}$ is assigned to a random variable, say x_v, with parents, say $x_{pa(v)}$, with a probability distribution:

$$p_v = p\left(x_v | x_{pa(v)}\right). \tag{1}$$

Assuming conditional independencies among the random variables, (1) can be re-written as:

$$p_v = \prod_{c \in \boldsymbol{C}} p\left(x_c | x_{pa(c)}\right) \prod_{d \in \boldsymbol{D}} p\left(x_d | x_{pa(c)}, x_{pa(d)}\right). \tag{2}$$

where, $p\left(x_d | x_{pa(c)}, x_{pa(d)}\right)$ is the conditional probability of x_c given the parents of both the discrete ($x_{pa(c)}$, set $\boldsymbol{C}$) and the continuous ($x_{pa(d)}$, set $\boldsymbol{D}$) variables, and $p\left(x_c | x_{pa(c)}\right)$ is the conditional probability of x_c given $x_{pa(c)}$. The DAG structure is used to generate new instances consistent with causal dependencies between the features. If the node is discrete, the probability distribution in (1) is uniform, otherwise a mean and a variance is attached per discrete parent configuration.

2) Tree ensembles

Both supervised tree ensembles (STE) and unsupervised tree ensembles (UTE) were used for virtual population generation [8-10]. In the supervised schema, an ensemble of decision trees is built, where in each tree node, the univariate empirical cumulative distribution function (ECDF) of the splitting feature is captured. In the unsupervised schema, a density forest ensemble is built, where the ensembles are density trees instead of decision trees [8-10]. In this case, the variance is used for the selection of the splitting feature.

3) Artificial neural networks

Artificial neural networks (ANNs) were also used for virtual population generation, where Gaussian radial basis functions (RBFs) are used as activation functions [9, 10]. In this case, the Gaussian RBFs are defined as in:

$$y(\boldsymbol{x}) = \sum_{i=1}^{N} w_i \, exp\left(-\beta ||\boldsymbol{x} - x_i||^2\right), \tag{3}$$

where $\boldsymbol{x}$ is an input vector with N-features, x_i is the center vector, $y(\boldsymbol{x})$ is the output, w_i is the weight of the i-th neuron, and β is a standard Gaussian parameter.

4) Gaussian Mixture Models with variational inference

A Gaussian mixture model (GMM) is a probabilistic model which assumes that the samples are generated from a mixture of a finite number of Gaussian distributions with unknown parameters [15]. A GMM approximation is defined as:

$$q(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{i=1}^{N} q(i; \boldsymbol{\theta}) q(\boldsymbol{x} | i; \boldsymbol{\theta}), \tag{4}$$

where i is the mixture component, $\boldsymbol{\theta}$ is the set of hyper-parameters, $q(i; \boldsymbol{\theta})$ are the mixture weights, and $q(\boldsymbol{x} | i; \boldsymbol{\theta})$ is a multivariate normal distribution (MVND) with mean $\boldsymbol{\mu_o}$ and covariance matrix $\boldsymbol{\Sigma_o}$, $N(\boldsymbol{x} | \boldsymbol{\mu_o}, \boldsymbol{\Sigma_o})$. A common approach for estimating $\boldsymbol{\theta}$ is based on the expectation-maximization algorithm which maximizes the data likelihood [15]. The EM, however, might yield GMMs with topologies that might not fit well to the underlying data structures. A solution to this is provided by variational inference (VI), which seeks for a lower bound on the model evidence instead of the likelihood. The goal of the GMM with variational Bayesian inference (BGMM) is to estimate the hyper-parameter(s) $\boldsymbol{\theta}$ in $q(\boldsymbol{x}; \boldsymbol{\theta})$, so that the Kullback-Leibler (KL) divergence with the posterior distribution $p(\boldsymbol{x})$ is minimized.

The KL-divergence is defined as in [16]:

$$KL\left(q(\boldsymbol{x}; \boldsymbol{\theta}) || p(\boldsymbol{x})\right) = \int_{\boldsymbol{x}} q(\boldsymbol{x}; \boldsymbol{\theta}) \, log\left(\frac{q(\boldsymbol{x}; \boldsymbol{\theta})}{p(\boldsymbol{x})}\right) d\boldsymbol{x}, \tag{5}$$

where the quotient of the search model over the posterior is the logarithm of the evidence, $L(\boldsymbol{\theta})$. Minimizing (5) is the same as maximizing a lower bound on $L(\boldsymbol{\theta})$:

$$argmax_{\theta}\left[\int_{\boldsymbol{x}} q(\boldsymbol{x}; \boldsymbol{\theta})\left(log\left(p(\boldsymbol{x})\right) - log(q(\boldsymbol{x}; \boldsymbol{\theta}))\right) d\boldsymbol{x}\right], \tag{6}$$

which refers to as the Evidence Lower Bound Objective (ELBO) [17]. In the case of GMM, where the search model is a multivariate normal distribution, (6) becomes:

$$argmax_{\theta}\left[\int_{\boldsymbol{x}} q(\boldsymbol{x}; \boldsymbol{\theta})\left(R(\boldsymbol{x}) + log\left(\tilde{q}(i|\boldsymbol{x})\right)\right) d\boldsymbol{x} + H(q)\right], \tag{7}$$

where $R(\boldsymbol{x})$ is equal to $log(p(x))$ and $H(q) = H\left(q(x|i)\right) = -\int_{\boldsymbol{x}} q(\boldsymbol{x}; \boldsymbol{\theta}) \, log\left(q(\boldsymbol{x}; \boldsymbol{\theta})\right) d\boldsymbol{x}$ is the entropy of $q(\boldsymbol{x}; \boldsymbol{\theta})$.

5) Proposed BGMM approach

Due to its Bayesian rationale, VI needs more hyper-parameters than EM, the most important of these being the concentration prior of the BGMM [17-19]. A common practice is to define the BGMM's prior structure according to a Dirichlet process mixture with concentration (or gamma) values equal to the inverse of the number of the Gaussian components [17-19]. In that case, however, a small weight concentration combined with many components would have a negative impact in the performance of the BGMM [17-19]. In this work, we utilize Dirichlet process mixtures as the BGMM's prior structure, where the concentration of each component on the weight distribution is defined as an exponential function of different Gaussian components to yield a stable number of components across multiple runs.

C. Virtual data quality evaluation

1) Variability and explainability

A key issue in virtual population generation lies in the underlying variability of the associations among the features in the virtual data which directly affects the explainability of the virtual data generators. In this work, we compute the intra-correlation as the average of the correlation differences between the real and the virtual data, on a feature basis, to examine whether the associations between the features in the virtual data are preserved across multiple runs. The inter-

correlation was also calculated as the overall mean correlation difference to examine the explainability of the generators.

2) Kolmogorov-Smirnov (KS) goodness-of-fit (gof) test

The goodness of fit (gof) test statistic [8] is used to quantify the similarity among the real and the virtual data as described in [8]. Large gof value denotes distributions with increased similarity with the absence of statistical significance.

3) Kullback-Leibler (KL) divergence

The Kullback-Leibler *(KL)* divergence [20] is defined as in (5), where $q(\boldsymbol{x};\boldsymbol{\theta})$ and $p(\boldsymbol{x})$ are replaced by the probability densities of the real data and the virtual data, respectively. KL values close to 0 denote distributions with small divergence.

III. Results

A. Data quality evaluation

All detected outliers and duplicated fields, as well as, features with high number of missing records were removed from further analysis. The final curated dataset included 11 features, namely the: (i) "Ech_Echo_LA" (Left Atrium), (ii) "Ech_Echo_LVIDs" (Left ventricular internal dimension), "ABNORMAL_HOLTER" (Abnormal Holter indicator), "Ech_Echo_Aortic_Root", "NYHA" (New York Heart Association class), "ARRHYTHMIA_NSVT" (Non sustained ventricular tachycardia), "Ech_Echo_PW" (Pulse Wave Doppler), "BMI" (Body Mass Index), "BSA" (Body Surface Area), "Height", "High_Risk". These features were used to evaluate the generators across multiple virtual patients in the range [1000, 20000] with a step 1000.

B. BGMM hyperparameter tuning

The average goodness of fit and inter-correlation values are depicted in Fig. 1 for components in the interval [1, 30]. For illustration purposes, the number of virtual patients has been restricted in the interval [1000, 10000].

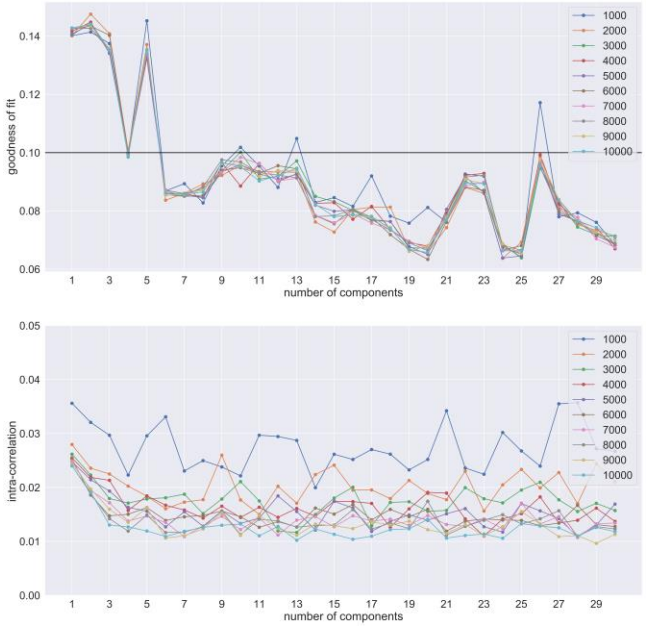

Figure 1: Performance evaluation of the proposed BGMM across multiple virtual patients in range [1000,10000].

According to Fig. 1, the average gof value was less than 0.1 for more than 5 Gaussian components. The average inter-correlation difference was less than 0.04 across the multiple virtual populations' executions and in some executions even less than 0.03. The average goodness of fit and correlation values from the four most prominent Gaussian components of Fig. 1 (i.e., for 19, 20, 24, and 25 components) are depicted in Fig. 2, along with the corresponding KL divergence and log-likelihood scores (which are referred to as BGMM scores). According to Fig. 2, the number of components that yielded virtual data with the smallest goodness of fit, KL divergence scores, correlation values, and the highest BGMM scores, across all executions, was 24. This number was combined with the Dirichlet concentration (gamma) value to generate multiple virtual patients.

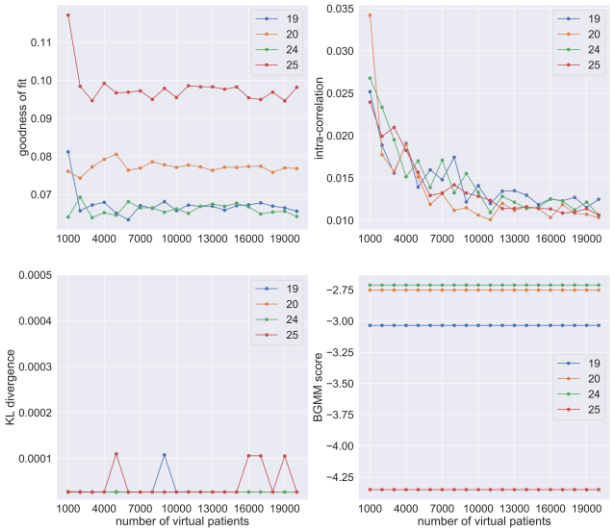

Figure 2: Performance evaluation of the proposed BGMM for the four best components across multiple virtual patients.

C. Performance comparison

For comparison purposes, the number of virtual patients was set to 20000. According to Table 1, the proposed BGMM approach achieved the lowest gof (less than 0.1) along with the UTE and the STE compared to the RBF-based ANN and the Bayesian networks. In addition, the proposed BGMM method yielded the lowest inter- and intra-correlation differences between the features in the virtual data (0.0133 inter-correlation and 0.0121 intra-correlation). In all cases, the average KL divergence was less than 0.001.

TABLE I: PERFORMANCE EVALUATION RESULTS.

Method	Average performance evaluation measures			
	Goodness of fit	Inter-correlation difference	Intra-correlation difference	KL divergence
BGMM	0.0667	0.0133	0.0121	<0.001
UTE	0.0211	0.0309	0.0281	<0.001
STE	0.0261	0.0433	0.0393	<0.001
ANN	0.1872	0.0829	0.0753	<0.001
Bayesian	0.1864	0.0824	0.0749	<0.001

According to Fig. 4, the average execution time of the proposed BGMM approach was faster than the UTE and the STE methods, yielding multiple virtual populations in 4.321 sec against the UTE and the STE which required 46.537, and 34.096 sec, respectively. The gap in the proposed BGMM during the generation of 10000 patients is related to the fast

convergence of the BGMM. The average execution times of the ANNs and the Bayesian methods were ignored due to their reduced performance against the previous methods.

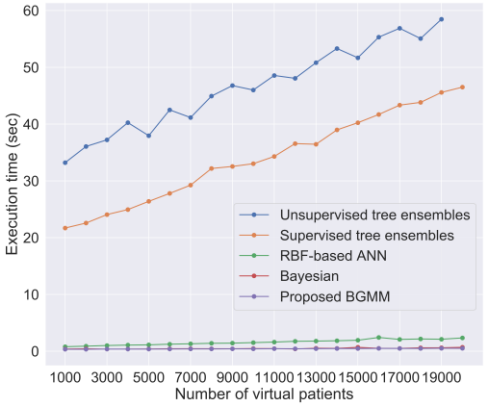

Figure 3: Execution time (sec) per virtual data generator.

IV. Conclusions

In this work, we utilized probabilistic Gaussian Mixture Models with variational Bayesian inference (BGMM) for the generation of large-scale virtual populations for *in-silico* clinical trials in HCM. The proposed approach uses weight concentration values for variational inference which are based on an exponentially decaying transformation of the number of Gaussian components. The proposed approach was compared against state-of-the-art virtual data generators, including, the Bayesian networks, the supervised tree ensembles (STE), the unsupervised tree ensembles (UTE), and the ANN yielding better inter- and intra- correlation differences in less execution time than the unsupervised tree ensembles which achieved the second-best performance.

The proposed method for the estimation of the Dirichlet concentration of each component on the weight distribution yielded a stable number of components (24 components) across multiple virtual populations executions, where the prior structure of the GMM was defined according to the Dirichlet process mixture. The proposed BGMM with the optimal number of Gaussian components achieved the lowest goodness of fit values (less than 0.1) along with the UTE and the STE compared to the RBF-based ANN and the Bayesian networks (with average gof larger than 0.15). In addition, the proposed BGMM method yielded the lowest inter- and intra-correlation differences between the features in the virtual data (almost 0.01), in less execution time (0.4319 sec) than the STE (46.5373 sec), which had the second-best performance. This confirms the computational efficiency of the proposed BGMM approach towards the generation of large-scale virtual populations for *in-silico* clinical trials in HCM.

As a future work, we plan to extend the proposed approach for data augmentation in other clinical domains, apart from *in-silico* clinical trials, as well as, to investigate the effect of the Dirichlet processes on the prior structure of the Gaussian Mixture Models to yield more robust finite mixture models.

References

[1] S. Sinisi, V. Alimguzhin, T. Mancini, E. Tronci, and B. Leeners, "Complete populations of virtual patients for *in silico* clinical trials," Bioinformatics, p. btaa1026, Dec. 2020.

[2] T. R. Rieger, R. J. Allen, L. Bystricky, Y. Chen, G. W. Colopy, Y. Cui, A. Gonzalez, Y. Liu, R. D. White, R. A. Everett, H. T. Banks, and C. J. Musante, "Improving the generation and selection of virtual populations in quantitative systems pharmacology models," Progress in biophysics and molecular biology, vol. 139, pp. 15-22, Nov. 2018.

[3] L. Zhao, M. J. Kim, L. Zhang, and R. Lionberger, "Generating model integrated evidence for generic drug development and assessment," Clin Pharmacol Therap, vol. 105, no 2, pp. 338-349, Feb. 2019.

[4] F. Stader, H. Kinvig, M. A. Penny, M. Battegay, M. Siccardi, and C. Marzolini, "Physiologically based pharmacokinetic modelling to identify pharmacokinetic parameters driving drug exposure changes in the elderly," Clinical pharmacokinetics, vol. 59, no 3, pp. 383-401, Mar. 2020.

[5] A. Tucker, Z. Wang, Y. Rotalinti, and P. Myles, P, "Generating high-fidelity synthetic patient data for assessing machine learning healthcare software," NPJ digital medicine, vol. 3, no 1, pp. 1-13, Nov. 2020.

[6] V. Pezoulas, N. Tachos, and D. Fotiadis, "Generation of virtual patients for in silico cardiomyopathies drug development," In Proceedings of the 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), pp. 671-674, Oct. 2019.

[7] M. Sood, A. Sahay, R. Karki, M. A. Emon, H. Vrooman, M. Hofmann-Apitius, and H. Fröhlich, "Realistic simulation of virtual multi-scale, multi-modal patient trajectories using Bayesian networks and sparse auto-encoders," Scientific reports, vol. 10, no. 1, pp. 1-14, Jul. 2020.

[8] V. C. Pezoulas, G. I. Grigoriadis, N. S. Tachos, F. Barlocco, and I. Olivotto, and D. I. Fotiadis, "Generation of virtual patient data for in-silico cardiomyopathies drug development using tree ensembles: a comparative study," In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), pp. 5343-5346, Jul. 2020.

[9] M. Robnik-Sikonja, "Package 'semiArtificial'", CRAN, 2019.

[10] M. Robnik-Šikonja, "Dataset comparison workflows," International Journal of Data Science, vol. 3, no 2, pp. 126-145, May 2018.

[11] V. Pinto, and R. Sooriyarachchi, "Comparison of methods of estimation for a goodness of fit test–an analytical and simulation study," Journal of Statistical Computation and Simulation, vol. 1-21, Jan. 2021.

[12] L. Velicki, D. G. Jakovljevic, A. Preveden, M. Golubovic, M. Bjelobrk, A. Ilic, S. Stojsic, F. Barlocco, M. Tafelmeier, N. Okwose, M. Tesic, P. Brennan, D. Popovic, G. A. MacGowan, A. Ristic, N. Filipovic, L. S. Maier, and I. Olivotto, "Genetic determinants of clinical phenotype in hypertrophic cardiomyopathy," BMC Cardiovascular Disorders, vol. 20, no 1, pp. 1-10, Dec. 2020.

[13] 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, and D. I. Fotiadis, "Medical data quality assessment: On the development of an automated framework for medical data curation," Computers in biology and medicine, vol. 107, pp. 270-283, Apr. 2019.

[14] M. Scanagatta, A. Salmerón, and F. Stella, "A survey on Bayesian network structure learning from data," Progress in Artificial Intelligence, vol. 8, no 4, pp. 425-439, May 2019.

[15] S. Meemansa, S. Akrishta, K. Reagon, E. M. Asif, V. Henri, M. Hofmann-Apitius, and F. Holger, "Realistic simulation of virtual multi-scale, multi-modal patient trajectories using Bayesian networks and sparse auto-encoders," Scientific Reports, vol. 10, no. 1, Jul. 2020.

[16] N. Manouchehri, H. Nguyen, P. Koochemeshkian, N. Bouguila, and W. Fan, "Online Variational learning of Dirichlet process mixtures of scaled Dirichlet distributions," Information Systems Frontiers, vol. 22, no. 5, pp. 1085-1093, Jul. 2020.

[17] V. Gallego, and D. Ríos Insua, "Variationally inferred sampling through a refined bound," Entropy, vol. 23, no 1, pp. 123, Feb. 2020.

[18] N. Manouchehri, H. Nguyen, P. Koochemeshkian, Bouguila, N., and W. Fan, "Online Variational learning of Dirichlet process mixtures of scaled Dirichlet distributions," Information Systems Frontiers, vol. 22, no. 5, pp. 1085-1093, Jul. 2020.

[19] H. Nguyen, M. Kalra, M. Azam, and N. Bouguila, "Data clustering using online variational learning of finite scaled dirichlet mixture models," In Proceedings of the 2019 IEEE 20th international conference on information reuse and integration for data science (IRI), pp. 267-274, Aug. 2019.

[20] S. Ji, Z. Zhang, S. Ying, L. Wang, X. Zhao, and Y. Gao, "Kullback-Leibler Divergence Metric Learning," IEEE Transactions on Cybernetics, pp. 1-12, Jul. 2020.