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

A federated AI strategy for the classification of patients with Mucosa Associated Lymphoma Tissue (MALT) lymphoma across multiple harmonized cohorts

Vasileios C. Pezoulas, *Student Member, IEEE,* Fanis Kalatzis, Themis P. Exarchos, *Member, IEEE,* Luke Chatzis, Saviana Gandolfo, Andreas Goules, Salvatore De Vita, Athanasios G. Tzioufas, and Dimitrios I. Fotiadis, *Fellow, IEEE*

Abstract—**Mucosa Associated Lymphoma Tissue (MALT) type is an extremely rare type of lymphoma which occurs in less than 3% of patients with primary Sjögren's Syndrome (pSS). No reported studies so far have been able to investigate risk factors for MALT development across multiple cohort databases with sufficient statistical power. Here, we present a generalized, federated AI (artificial intelligence) strategy which enables the training of AI algorithms across multiple harmonized databases. A case study is conducted towards the development of MALT classification models across 17 databases on pSS. Advanced AI algorithms were developed, including federated Multinomial Naïve Bayes (FMNB), federated gradient boosting trees (FGBT), FGBT with dropouts (FDART), and the federated Multilayer Perceptron (FMLP). The FDART with dropout rate 0.3 achieved the best performance with sensitivity 0.812, and specificity 0.829, yielding 8 biomarkers as prominent for MALT development.**

Keywords: federated AI, data harmonization, Mucosa Associated Lymphoma Tissue (MALT), primary Sjögren's Syndrome (pSS).

I. Introduction

Primary Sjögren's Syndrome (pSS) is mainly characterized by dry eye and mouth manifestations with an increased risk of evolution into malignant lymphoma [1-3]. It is a chronic systemic autoimmune disease with diverse clinical picture and outcome [1-3] and is primarily affecting middle-aged women [1-3]. The estimated prevalence of pSS ranges from 0.01-0.1% of the general population and it is affected by the geographic distribution and the classification criteria [1-3]. It has a wide range of clinical presentations, extending from mild disease limited to exocrine glands to severe, multi-systemic disorder, and development of B-cell non-Hodgkin lymphoma (NHL) in about 5% of patients [1-3].

The clinical unmet needs in pSS include the development of lymphoma classification and lymphomagenesis models, as well as, the extraction of prominent indicators for lymphoma development, as potential biomarkers. In the majority of these studies, both univariate and multivariate statistical models [4-6], as well as, time-to-event models [6] have been employed to detect risk factors. In another study [7], gradient boosting trees were used for the first time in pSS to predict lymphoma outcomes in a single cohort of 435 patients yielding increased accuracy and sensitivity, as well as, on four European cohorts (1554 patients) [8] with notable performance.

All the above studies mainly focus on the development of lymphoma classifiers for general types of lymphoma (e.g., by grouping patients with Mucosa Associated Lymphoma Tissue (MALT) and patients with Diffuse Large B-cell Lymphoma (DLBCL)) [9-11] and, thus, conceal significant clinical evidence regarding rare types of lymphoma in pSS, such as, MALT, which occurs in less than 3% of pSS patients. In addition, most of the studies focus on the application of conventional multivariate methods for the extraction of independent risk factors for lymphoma development without giving any particular emphasis on the performance of the resulting models [9-11]. Moreover, the existing studies focus either on training robust classifiers, such as, tree ensembles on a single cohort of pSS patients [7] or on small cohorts [8], and, thus, obscure the statistical power of the outcomes due to the lack of sufficient population size in such a rare disease.

To address these needs, we have developed a federated AI framework that enables the training of trustworthy AI algorithms across cloud databases, which includes the: (i) federated gradient boosting trees (FGBT), (ii) FGBT with dropouts (FDART), (iii) federated Multilayer Perceptron (FMLP), and (iv) federated Multinomial Naïve Bayes (FMNB). A large-scale case study was conducted towards the development of AI models for MALT classification across 17 curated and harmonized cohorts. The FDART with dropout rate 0.3 achieved the best performance with accuracy 0.828, sensitivity 0.812, and specificity 0.829, where the AI model's explainability was validated by the detection of 8 biomarkers for MALT. To our knowledge, this is the first case study that develops federated MALT classification models in pSS.

*This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreemend No. 739144 and from the Swiss State Secretariat for Education, Research and Innovation (SERI), under grant agreement No. 16.0210.

V.C. Pezoulas, F. Kalatzis, and D.I. Fotiadis 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, tkalatz@gmail.com, and fotiadis@uoi.gr).

T.P. Exarchos is with the Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, University of Ioannina, GR 45110 Ioannina, Greece, and with the Dept. of Informatics, Ionian University, Corfu GR 49100, Greece (e-mail: exarchos@cc.uoi.gr).

S. Gandolfo, and S. De Vita are with the Clinic of Rheumatology, Department of Medical and Biological Sciences, Udine University, Udine, IT33100, Italy (e-mails: saviana.gandolfo@uniud.it, salvatore.devita@asuiud.sanita.fvg.it).

L. Chatzis, A. Goules, and A.G. Tzioufas are with the Dept. of Pathophysiology, Faculty of Medicine, National and Kapodistrian University of Athens, GR 15772 Athens, Greece (emails: lukechatzis@gmail.com, agoules@med.uoa.gr, agtzi@med.uoa.gr).

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

II. Materials And Methods

A. Data sharing

Pseudonymized patient data were collected from 21 European cohorts with 7,551 patients who have been diagnosed with primary Sjogren's Syndrome (pSS) [1]. The cohort data were obtained under a data protection agreement, fulfilling all the necessary ethical and legal requirements for data sharing that are posed by the General Data Protection Regulation. Upon the GDPR approval, the pseudonymized patient data were stored in secure federated cloud databases.

Only those databases having at least one patient who has been diagnosed with any type of lymphoma were included in the analysis. According to Fig. 1, the 17 remaining cohorts were ranked in descending order based on the number of MALT patients. Cohorts UiB, MHH and CUMB included lymphoma patients but none of them had MALT lymphoma type and thus were used to populate the control group.

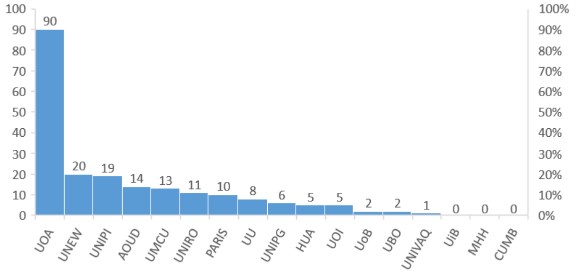

Figure 1. Distribution of pSS patients with MALT lymphoma type across the 17 harmonized cloud databases.

B. Data curation and harmonization

A clinical data curation pipeline presented in a previous study [12] was applied to resolve data inconsistencies, such as, outliers, missing values, and highly correlated features. The pipeline was applied on each individual raw database. Both univariate and multivariate methods, such as, the z-score and the interquartile range along with the density-based methods [12, 13] were used to isolate anomalies. Correlation matrices were constructed to detect highly correlated features as potential duplicates [13]. Data harmonization was applied according to a reference pSS ontology which was presented in a previous study [14]. The reference ontology consists of a set of clinical parameters which reflect the minimum available domain knowledge of pSS. The latter was used to semi-automatically align the terminologies of each dataset with the reference ontology upon clinical guidance [8].

C. Federated AI framework

Since the harmonized cohort data were stored in private cloud databases, the data integration approach was not feasible, thus obscuring the training of conventional machine learning (ML) algorithms. To deal with this challenge, we designed a federated AI framework which focuses on the incremental training of ML models across federated databases. The overall process is depicted in Fig. 2. A central computing node (CCN) is used to incrementally connect and communicate with the federated databases $\boldsymbol{D_1}, \boldsymbol{D_2}, \ldots, \boldsymbol{D_N}$. On each training phase, the ML models $\boldsymbol{M_1}, \boldsymbol{M_2}, \ldots, \boldsymbol{M_N}$ were incrementally updated across the databases, in continuous time steps, and evaluated either on one or on more databases.

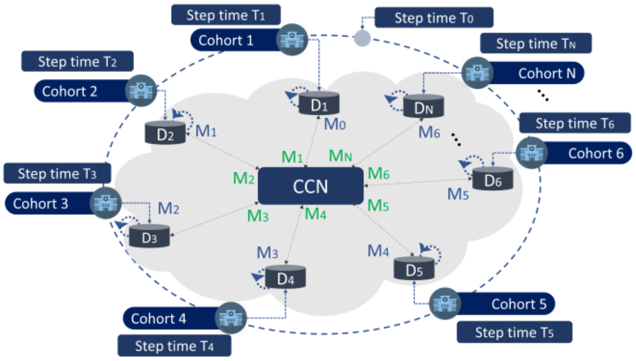

Figure 2. An illustration of the federated AI framework.

1) Rationale

For a given a set of N-databases, say $\{\boldsymbol{D_1}, \boldsymbol{D_2}, \ldots, \boldsymbol{D_N}\}$, a machine learning algorithm trained on the dataset $\boldsymbol{d_i} \in \boldsymbol{D_i}$ is updated through the following function [13]:

$$F(\boldsymbol{d_i}) = F(\boldsymbol{d_i} - 1) + \beta q(\boldsymbol{d_i}), \quad (1)$$

where, $F(\boldsymbol{d_i})$ corresponds to the estimated ML model which has been trained on the dataset $\boldsymbol{d_i}$, $F(\boldsymbol{d_i} - 1)$ corresponds to the estimated model which was trained on $\boldsymbol{d_{i-1}}$, $q(\boldsymbol{d_i})$ is the learner on $\boldsymbol{d_j}$, and β is a scalar. A loss function can then be defined in the form $L(f(\boldsymbol{d_i}), \boldsymbol{y_i})$ where $f(\boldsymbol{d_i})$ is the estimator and $\boldsymbol{y_i}$ is the target score. Then, the stochastic gradient descent (SGD) approach is used to minimize the loss function through the following sequential weight update process [15]:

$$w(\boldsymbol{d_i}) = w(\boldsymbol{d_i} - 1) - \beta(\nabla_w L(f(\boldsymbol{d_i}), \boldsymbol{y_i}) + a\nabla_w r(w)), \quad (2)$$

where, $\nabla_w L(f(\boldsymbol{d_i}), y_i)$ is the gradient of the loss function with respect to w, $r(w)$ is a regularization function, $\nabla_w r(w)$ is the gradient of the regularization function, a is a hyperparameter, and β is a learning rate parameter.

2) Stochastic gradient descent and probabilistic methods

All supervised learning classifiers which adopt the SGD function can be extended through (2) to support incremental learning. If we replace the loss function with the hinge loss:

$$\varphi(f(\boldsymbol{x_i}), \boldsymbol{y_i}) = max\,(0, -\boldsymbol{y_i} f(\boldsymbol{x_i})), \quad (3)$$

we construct the Perceptron classifier. In a similar way, we can obtain federated artificial neural network classifiers, such as, the Multi-layer perceptron (FMLP) [15]. As for the probabilistic methods, the federated Multinomial Naïve Bayes (FMNB) is used as in [15].

3) Federated gradient boosting trees

In the case of federated gradient boosting trees, the goal is to obtain an AI model using ensembles of regression trees, as weak learners, which minimize the expected value of the loss function. Gradient boosting trees were used by incrementally seeking for the mapper $F(\boldsymbol{x})$ at a stage m, $F_m(x)$, as in [14]:

$$F_m(\boldsymbol{d_i}) = F_{m-1}(\boldsymbol{d_i}) + p_m \cdot h(\boldsymbol{d_i}; \boldsymbol{a_m}), \quad (4)$$

where p_m is the line search, and $h(\boldsymbol{x}; \boldsymbol{a_m})$ is a regression tree learner with parameters $\boldsymbol{a_m}$. A main issue in gradient boosting trees is the fact that the algorithm combines many regression trees with a small learning rate and thus trees that are added early in the ensemble are more significant than trees added late. To deal with this issue, we used the DART approach [16] according to which the dropped trees and the new tree, on each round, were scaled by a factor which ensures that the combination of the dropped trees with the new trees will have the same effect on the outcome [16]. The DART is trained on intermediate datasets of the random subset that is selected by the gradient boosting trees and thus prevents the construction of trivial trees. For a given model, say M, with $M(t)$ denoting the prediction for point x, DART creates the subset [16]:

$$\left\{\left(t, -\nabla_t L\left(T(t)\right)\right)\right\}, \tag{5}$$

where $\nabla L\left(T(t)\right)$ is the gradient of the loss function $L\left(T(t)\right)$. Thus, a new label with values $-\nabla_t L\left(T(t)\right)$ is created for each sample t in the training dataset.

III. Results

A. Federated data storage

Each data provider uploaded his/her cohort data in secure federated cloud databases within the HarmonicSS platform [17]. The legal and ethical compliance of the cohort data was evaluated by a Data Controllers Committee (DCC) which consisted of one technical and two clinical experts in the field. Upon the DCC approval, the cohort data curation and harmonization workflows were applied on the raw data.

B. High-quality and harmonized federated databases

All detected outliers and data incompatibilities, as well as, duplicated fields were automatically discarded from each federated database. The curated data were utilized in the data harmonization pipeline using the pSS reference ontology as a gold standard which yielded 4,805 harmonized patients (206 MALT, 4,599 non-MALT). The harmonized data included 42 features with more than 90% overlap with the pSS reference ontology which were used to train federated AI algorithms for MALT classification (0: absence, 1: presence).

C. Federated MALT classification

To control for the increased class imbalance between the two populations, we applied random downsampling with replacement on each training cohort, where the ratio between the MALT and non-MALT patients was set to 1:1. On each random subset, the patients were matched according to age at SS diagnosis, gender, and disease duration, yielding 4805 patients, in total. Since all the possible permutations of the training databases is 17!, the databases were ranked in descending order according to the number of MALT patients, as in Fig. 1. Databases having less than 10 MALT patients were excluded from the permutations thus restricting the number of possible training permutations to 7!. For illustration purposes, the analysis was restricted in two cases, where the training sequence included the following harmonized cohort databases: UOA, UNEW, UNIPI, UMCU, UNIRO, PARIS, UU, HUA, UOI, UoB, UBO, UNIVAQ. The validation was conducted in two testing cohorts: AOUD (case 1) and UNIPG (case 2).

According to Table 1, the FGBT and FDART algorithms achieved the best performance, where the FDART with dropout rate (rd) 0.3 had the best performance (accuracy = 0.828, sensitivity = 0.812, specificity = 0.829, and AUC = 0.853). The FMLP and the FMNB achieved the lowest performance due to overfitting effects which were introduced in the weight update process during the training in cohorts having reduced number of MALT patients.

TABLE I. Performance Evaluation Results Case 1.

Algorithm	Accuracy	Sensitivity	Specificity	AUC
Case 1 (Starting cohort: UOA, testing cohort: AOUD)				
FGBT	0.808	0.625	0.818	0.84
FDART, rd = 0.1	0.805	0.812	0.804	0.825
FDART, rd = 0.2	0.795	0.812	0.794	0.865
FDART, rd = 0.3	0.828	0.812	0.829	0.853
FDART, rd = 0.4	0.818	0.812	0.818	0.852
FMNB	0.502	0.875	0.48	0.678

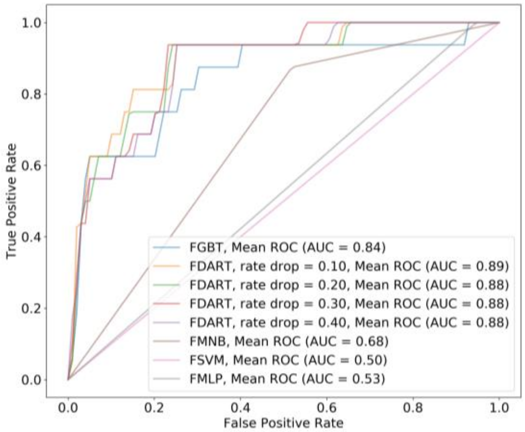

Figure 3. Receiver Operating Characteristic (ROC) curves for MALT classification using the FDART for different dropout rates against the federated FGBT, FMNB, FSVM, and FANN (case 1).

In the second case, the FGBT and the FDART algorithms achieved once more the best performance but with a reduced quality than in the previous case due to the small percentage of MALT patients in the UNIPG cohort (3.6% ratio with 6 MALT cases) compared to the AOUD cohort (5% ratio with 14 MALT cases). Once again, the FMLP and the FMNB had the lowest performance due to additional effects which were introduced during the training phase.

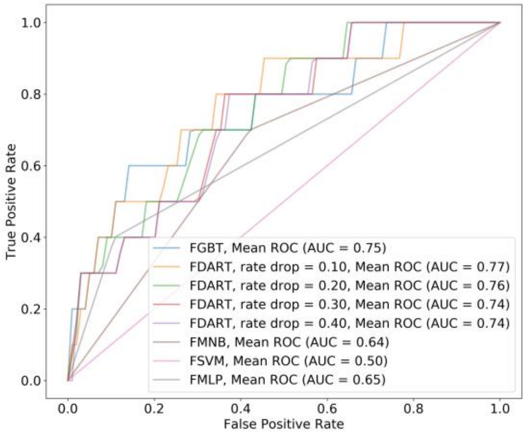

Figure 4. Receiver Operating Characteristic (ROC) curves for MALT classification using the FDART for different dropout rates against the federated FGBT, FMNB, FSVM, and FMLP (case 2).

D. Explainability of the federated MALT classifiers

To further investigate the explainability of the federated AI models, we have induced the instances of the tree ensembles that have highly participated in the decision-making process.

In both cases, a set of prominent features (biomarkers) was extracted using the F-score as a measure of the frequency of each feature in the node splitting process on each tree [18]. The set of biomarkers includes the: (i) Low C4 (F-score 12.6), (ii) Rheumatoid Factor (F-score 11.8), (iii) Cryoglobulinemia (F-score 10.8), (iv) Parotid or Submandibular swelling (F-score 10.4), (v) Arthritis (F-score 6.4), (vi) Raynaud's phenomenon (F-score 3.6), (vii) Palpable Purpura (F-score 3.4), and (viii) Renal disease (F-score 1.4), among others.

IV. Conclusions

In this work, we presented a generalized, federated AI framework which was built on the HarmonicSS [17] cloud infrastructure to enable the development of federated AI models across federated cloud databases with patients who have been diagnosed with pSS. The federated AI framework was utilized to train tree ensembles (federated gradient boosting trees with and without dropouts), stochastic gradient descent methods (federated Multilayer Perceptron), as well as, probabilistic methods (federated Multinomial Naïve Bayes) for MALT classification across 17 harmonized cloud databases. A case study was performed, where a federated AI model was trained on 15 cohorts and tested on 2 cohorts.

The FDART with dropout rate 0.3 (accuracy = 0.828, sensitivity = 0.812, specificity = 0.829, and AUC = 0.853) achieved the best performance along with the FGBT (accuracy = 0.808, sensitivity = 0.625, specificity = 0.818, AUC = 0.84). The FMLP and the FMNB had poor performance since their decision boundaries were affected by weight overfitting during the application of (2) in cohorts with small number of MALT patients. This issue, however, was not present on the FDART, where decision-making is based on an ensemble of trees with adequate performance during training. The performance of the federated tree ensembles was also better in the case where the testing cohort was UNIPG but with a smaller performance than in the AOUD due to the small percentage of MALT patients in UNIPG. The explainability of the AI models was validated through the extraction of 8 biomarkers for MALT development, where the "Low C4", "Rheumatoid Factor", and "Raynaud's phenomenon" were reported in the literature [3, 7, 19, 20].

This is the first case study regarding the development of AI models for MALT classifiers across multiple harmonized cohort databases in pSS, where MALT is one of the rarest types of lymphoma with a frequency less than 3% in pSS patients. As a future work, we are planning to apply the federated AI framework for the development of robust AI models with more outcomes, such as, peripheral nervous system disease, as well as, across other types of lymphoma, such as, Diffuse Large B-cell Lymphoma (DLBCL), and Nodal Marginal Zone Lymphoma (NMZL), among others.

References

[1] A. V. Goules, and A. G. Tzioufas, "Lymphomagenesis in Sjögren's syndrome: predictive biomarkers towards precision medicine," Autoimmunity reviews, vol. 18, no 2, pp. 137-143, Feb. 2019.

[2] A. Travaglino, C. Giordano, M. Pace, S. Varricchio, M. Picardi, F. Pane, and M. Mascolo, "Sjögren syndrome in primary salivary gland lymphoma: a systematic review and meta-analysis," American journal of clinical pathology, vol. 153, no 6, pp. 719-724, Feb. 2020.

[3] E. K. Kapsogeorgou, M. Voulgarelis, and A. G. Tzioufas, "Predictive markers of lymphomagenesis in Sjögren's syndrome: from clinical data to molecular stratification," Journal of autoimmunity, vol. 104, pp. 102316, Nov. 2019.

[4] G. Nocturne, A. Virone, W. F. Ng, V. Le Guern, E. Hachulla, D. Cornec, et al, "Rheumatoid factor and disease activity are independent predictors of lymphoma in primary Sjögren's syndrome," Arthritis & rheumatology, vol. 68, no 4, pp. 977-985, Apr. 2016.

[5] G. Ingravallo, E. Maiorano, M. Moschetta, L. Limongelli, M. G. Mastropasqua, G. F. Agazzino, et al, "Primary Breast Extranodal Marginal Zone Lymphoma in Primary Sjögren Syndrome: Case Presentation and Relevant Literature," Journal of Clinical Medicine, vol. 9, no 12, pp. 3997, Nov. 2020.

[6] A. Igoe, S. Merjanah, and Scofield, R. H, "Sjögren Syndrome and Cancer," Rheumatic Disease Clinics, vol. 46, no 3, pp. 513-532, 2020.

[7] V. C. Pezoulas, T. P. Exarchos, A. G. Tzioufas, S. De Vita, and D. I. Fotiadis, "Predicting lymphoma outcomes and risk factors in patients with primary Sjögren's Syndrome using gradient boosting tree ensembles," In Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2165-2168, Oct. 2019.

[8] V. C. Pezoulas, K. D. Kourou, F. Kalatzis, T. P. Exarchos, E. Zampeli, S. Gandolfo, A. Goules, C. Baldini, F. Skopouli, S. De Vita, A. G. Tzioufas, and 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 (OJEMB), vol. 1, pp. 83-90, Mar. 2020.

[9] S. Retamozo, P. Brito-Zerón, and M. Ramos-Casals, "Prognostic markers of lymphoma development in primary Sjögren syndrome," Lupus, vol. 28, no 8, pp. 923-936, June 2019.

[10] A. V. Goules, O. D. Argyropoulou, V. C. Pezoulas, L. Chatzis, E. Critselis, S. Gandolfo, F. Ferro, M. Binutti, V. Donati, S. Zandonella Callegher, A. Venetsanopoulou, E. Zampeli, M. Mavrommati, P. V. Voulgari, T. Exarchos, C. P. Mavragani, C. Baldini, F. N. Skopouli, D. I. Fotiadis, S. De Vita, H. M. Moutsopoulos, A. G. Tzioufas, "Primary Sjögren's Syndrome of Early and Late Onset: Distinct Clinical Phenotypes and Lymphoma Development," Frontiers in immunology, vol. 11, pp. 2707, Oct. 2020.

[11] S. Fragkioudaki, C. P. Mavragani, and H. M. Moutsopoulos, "Predicting the risk for lymphoma development in Sjogren syndrome: an easy tool for clinical use," Medicine, vol. 95, no 25, June 2016.

[12] 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," CBM, vol. 107, pp. 270-283, Apr. 2019.

[13] V. C. Pezoulas, T. Exarchos, and D. I. Fotiadis, "Medical data sharing, harmonization and analytics," Academic Press, Elsevier, Jan. 2020.

[14] V. C. Pezoulas, T. P. Exarchos, V. Andronikou, T. Varvarigou, A. Tzioufas, S. De Vita, and D. I. Fotiadis, "Towards the establishment of a biomedical ontology for the primary Sjögren's Syndrome," in *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, pp. 4089-4092, Jul. 2018.

[15] V. C. Pezoulas, F. Kalatzis, T. P. Exarchos, A. Goules, S. Gandolfo, E. Zampeli, F. Skopouli, S. De Vita, A. G. Tzioufas, and D. I. Fotiadis, "Dealing with Open Issues and Unmet Needs in Healthcare Through Ontology Matching and Federated Learning," In Proceedings of EMBEC, vol. 80, pp. 306-313, Nov. 2020.

[16] R. K. Vinayak, and R. Gilad-Bachrach, "Dart: Dropouts meet multiple additive regression trees," In Artificial Intelligence and Statistics, pp. 489-497, Feb. 2015.

[17] A. V. Goules, T. P. Exarchos, V. C. Pezoulas, K. D. Kourou, A. I. Venetsanopoulou, S. De Vita, D. I. Fotiadis, and A. G. Tzioufas, "Sjögren's syndrome towards precision medicine: the challenge of harmonisation and integration of cohorts," Clin Exp Rheumatol, vol. 37, no. Suppl 118, pp. S175-84, Jul. 2019.

[18] T. Chen, and C. Guestrin, "Xgboost: A scalable tree boosting system," In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, Aug. 2016.

[19] S. De Vita, and S. Gandolfo, "Predicting lymphoma development in patients with Sjögren's syndrome," Expert review of clinical immunology, vol. 15, no 9, pp. 929-938, Sep. 2019.

[20] C. Skarlis, E. Argyriou, and C. P. Mavragani, "Lymphoma in Sjögren's Syndrome: Predictors and Therapeutic Options," Current Treatment Options in Rheumatology, vol. 6, no. 1, pp. 1-17, Jan. 2020.