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

Multiple additive regression trees with hybrid loss for classification tasks across heterogeneous clinical data in distributed environments: a case study

Vasileios C. Pezoulas, *Student Member, IEEE,* Themis P. Exarchos, *Member, IEEE,* Athanasios G. Tzioufas, and Dimitrios I. Fotiadis, *Fellow, IEEE*

***Abstract*— Multiple additive regression trees (MART) have been widely used in the literature for various classification tasks. However, the overfitting effects of MART across heterogeneous and highly imbalanced big data structures within distributed environments has not yet been investigated. In this work, we utilize distributed MART with hybrid loss to resolve overfitting effects during the training of disease classification models in a case study with 10 heterogeneous and distributed clinical datasets. Lexical and semantic analysis methods were utilized to match heterogeneous terminologies with 80% overlap. Data augmentation was used to resolve class imbalance yielding virtual data with goodness of fit 0.01 and correlation difference 0.02. Our results highlight the favorable performance of the proposed distributed MART on the augmented data with an average increase by 7.3% in the accuracy, 6.8% in sensitivity, 10.4% in specificity, for a specific loss function topology.**

Keywords: distributed environments, data augmentation, lexical analysis, multiple additive regression trees, hybrid loss.

I. Introduction

Nowadays, big data in healthcare can provide broader and more comprehensive insight on the clinical decision-making process, as well as, enhance the statistical power of the clinical research studies [1-3]. The most common strategy for knowledge distillation across complex big data structures is based on the integrative analysis of data from multiple clinical registries which are shared and stored in centralized databases [4]. This, however, is not always feasible either due to GDPR (General Data Protection Regulation) violations or due to computational burdens which arise during the analysis of big data [5]. A solution to this is to use distributed environments [6, 7], where the data are stored in distributed nodes.

A technical challenge in distributed environments is to train machine learning algorithms across clinical data which are distributed across multiple computing nodes [8]. Towards this direction, batch processing methods, such as, online learning and meta-learning [9, 10] have been proposed, where the former [9] uses stochastic optimization to update an existing estimator on upcoming training instances whereas the latter [10] focuses on the aggregation of outcomes from models which are trained on each distributed node. Meta-learning methods, however, limit the "horizon" of the training process since the individual models are trained on individual subsets [2] and are restricted to the additive update of the weights of the model on new "online" training instances. A solution to this is to use incremental learning [11-13] which trains a classifier on an initial dataset, and then incrementally adjusts its weights on a series of existing datasets. Towards this direction, several incremental learning algorithms have been proposed in [13] including the multiple additive regression trees (MART), the Support Vector Machines (SVM), and the Multinomial Naïve Bayes (MNB), among others, where the gradient boosting trees (a specific type of MART) have the best performance in many classification tasks [12, 13].

A common problem with multiple additive regression trees, however, is the fact that trees added early in the ensemble tend to have a higher impact during the decision-making process than the trees added later [14]. Dropouts have been recently adopted by the deep learning community [14] to deal with this issue by scaling the most prominent trees in the ensemble with a specific rate of rejected trees. However, a main problem in MART with dropout rates is to account for overfitting effects in the selection of the dropout rate which is arbitrary. Although the MART (with and without dropouts) have been widely used in the literature as robust classifiers for different disease classification tasks [13-15] none of these algorithms have investigated the overfitting effects of MART during the training across heterogeneous and highly imbalanced big data structures in distributed environments.

To deal with this issue, we propose a pipeline which utilizes distributed multiple additive regression trees (MART) with hybrid loss for training across 15 distributed and harmonized datasets of patients with autoimmune diseases. Data pre-processing routines were utilized for data quality control. Flexible and stringent data harmonization methods were developed to detect overlapping terminologies across the heterogeneous data. Density forest ensembles were used for data augmentation to yield high-quality virtual data. Our results highlight the performance of the proposed pipeline on the augmented data yielding an average increase by 6.8% in sensitivity, and 10.4% in specificity, for a specific loss function topology, compared to training only on the real data.

* The research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the HFRI PhD Fellowship grant (Fellowship Number: 1357).

V.C. Pezoulas, 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, 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).

A.G. Tzioufas is with the Dept. of Pathophysiology, Faculty of Medicine, National and Kapodistrian University of Athens, GR 15772 Athens, Greece (email: agtzi@med.uoa.gr).

II. Materials And Methods

A. Distributed data analytics pipeline

The pipeline that was developed in this study consists of three layers, namely the: (i) data-preprocessing layer, (ii) the data harmonization layer, and (iii) the distributed data analytics layer. The first layer focuses on data quality analysis through the elimination of data recording errors, such as, features with joint variability. The data harmonization layer resolves structural heterogeneities across distributed datasets through flexible and stringent lexical analysis. The final layer utilizes advanced ML algorithms for additive training across augmented datasets in distributed environments, where the integrative analysis is restricted either due to GDPR or due to the increased need of computational resources for analysis.

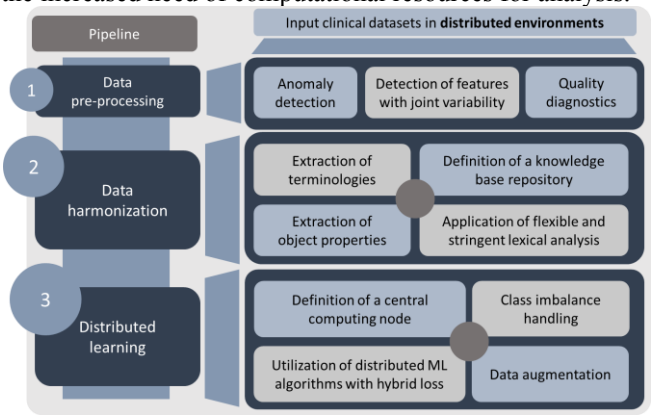

Figure 1. An illustration of the distributed analysis pipeline.

B. Data pre-processing

A data curation workflow presented in a previous study [16] was enhanced to support big data structures. Robust methods for anomaly detection were developed, including the elliptic gaussian curves [2] and the isolation forests [2], where the former detects anomalies by fitting multivariate Gaussian distributions and the latter trains tree ensembles to isolate outliers. The covariance and the Spearman correlation [2] were used to detect features with joint variability in the data.

C. Data harmonization

An automated harmonization workflow was utilized on top of a knowledge base repository which communicates with the OHDSI Athena vocabulary [17]. Two types of lexical analysis algorithms were developed, the "stringent" which detects lexical matches with high coherence and the "flexible" which proposes more matches with less coherence. Both algorithms solve the edit distance problem [2] using the Hamming and the Levenshtein distances [2] to detect terminologies with common string sequences. Optional semantic information was extracted from semantic data models that describe a domain knowledge using entities and object properties [18].

D. Machine learning in distributed environments

1) Strategy

Distributed learning lies on the additive adjustment of a single estimator across multiple data structures [8]. To achieve this we update the weights of the estimator through the stochastic gradient descent (SGD) method which seeks for a linear loss function, $h(f(x_i), y_i)$, which minimizes [11-13]:

$$L(\boldsymbol{w}) = argmin\left(\frac{1}{N}\sum_{i=1}^{N} h(f(x_i), y_i) + ar(\boldsymbol{w})\right), \quad (1)$$

where, x_i is the i-th instance, y_i is the target, $\boldsymbol{w}$ is a weight vector, $h(.)$ is a loss function, a is a hyperparameter, $r(\boldsymbol{w})$ is a regularizer, $L(.)$ is the objective, and $f(x_i)$ is a linear score function. Solving (1) yields the weight update formula:

$$w_i = w_{i-1} - \eta_t\left(\nabla_w h(f(x_i), y_i) + a\nabla_w r(w)\right), \quad (2)$$

where, i is the stage, w_{i-1} is the weight estimation at stage $i-1$, η_t is a non-negative learning rate parameter, and $\nabla_w h(f(\boldsymbol{x_i}), \boldsymbol{y_i})$ is the gradient of the loss function $h(.)$.

2) Distributed multiple additive regression trees (MART)

In the case of boosting ensembles (or multiple additive regression trees – MART), we seek for an estimator of week regression tree learners, at a training stage, i, as in [2, 11-13]:

$$F_i(\boldsymbol{x}) = F_{i-1}(\boldsymbol{x}) - \gamma_i \sum_{j=1}^{n} \nabla_{F_{i-1}} \varphi(y_j, F_{i-1}(x_j)), \quad (3)$$

where the regularization term in (3) is defined as in:

$$E(t) \approx \sum_{i=1}^{N}\left[\varphi(y_i, \tilde{y}_{i,t-1}) + g_i f_t(x_i) + \frac{1}{2} h_i {f_t}^2(x_i)\right], \quad (4)$$

with g_i denoting the first order gradient and h_i the second order gradient of the loss function. An underlying problem in MART is the fact that trees which are added early in the ensemble tend to have a higher impact in decision making than those added after [14]. A solution is to use a dropout rate [14], where the dropped trees on each stage are combined with the remaining trees through a scaling factor [14].

3) Distributed MART with hybrid loss

A main problem in distributed MART with dropout rates is to account for overfitting effects in the selection of the dropout rate, say r, we propose a hybrid loss function which combines the *logcosh* loss [19], say f, with the Huber loss, say g [20], where the loss topology is controlled by a δ value. An exponential function was defined between r and δ so that the shape of the loss function is steeper around 0 to avoid overfitting for large r. The first and second order gradients of f and g were used in (4), where the former are defined as in:

$$\nabla f = tanh(x), \qquad \nabla g = d/\sqrt{s}, \quad (5)$$

and the second order gradients are defined as in:

$$\nabla^2 f = 1/cosh^2(x), \qquad \nabla^2 g = \sqrt{s}/s, \quad (6)$$

where d and s are approximation factors [22]. Eqs (5), (6) are combined based on the product rule and utilized in Eq (4).

A pseudocode that summarizes the backbone of distributed learning is presented in Algorithm 1. An ML algorithm is trained on the first dataset yielding the initial weights which are additively updated across the rest of the datasets through (2) based on the weights from the previous executions.

Algorithm 1. A pseudocode for distributed learning.	
1	**def distributed_learning(F, $\boldsymbol{T} = \{\boldsymbol{T_0}, \boldsymbol{T_1}, \boldsymbol{T_2} \ldots, \boldsymbol{T_M}\}, \boldsymbol{w_0}$):**
2	fit an estimator $\boldsymbol{F_o}$ on the dataset $\boldsymbol{T_o}$ yielding $\boldsymbol{w_0}$
3	for $i = 0: M$ do:
4	retrieve weight vector w_{i-1} from the previous execution
5	solve $w_i = w_{i-1} - \eta_t\left(\nabla_w h(f(x_i), y_i) + a\nabla_w r(w)\right)$
6	update $F_i(\boldsymbol{x}) = F_{i-1}(\boldsymbol{x}) - \gamma_i \sum_{j=1}^{n} \nabla_{F_{i-1}} \varphi(y_j, F_{i-1}(x_j))$
7	return $[w_i, F_i]$

4) Data augmentation using density forest ensembles

Density forest ensembles were used as high-quality virtual data generators [21] instead of the conventional probabilistic methods which are restricted to oversampling with biased assumptions. Density trees are built in a top-down way, where the splitting process is based on the variance of each feature. A density forest is as a mixture of Gaussian densities [22]:

$$p(v) = \frac{1}{M}\sum_{k=1}^{M} p_k(v) = \frac{1}{M}\sum_{k,q} g_q(v) N\left(v; \mu_q(v), \Sigma_q(v)\right), \quad (7)$$

where $v \in V$ is a tree node, $N(v; \mu_q(v), \Sigma_q(v))$ is a multivariate Gaussian distribution with mean $\mu_q(v)$ equal to the mean of all points reaching the leaf $q \in Q$, $\Sigma_q(v)$ is the covariance and $g_q(v)$ is the proportion of all points reaching q. Statistical measures, such as, the goodness of fit (gof), the Kullback-Leibler (KL) divergence and correlation [21] were used to quantify the similarity among the real and virtual data.

III. Results

A. Data quality

Anonymized clinical data were collected from 10 databases with patients who have been diagnosed with primary Sjögren's Syndrome (pSS) under the HarmonicSS Project [23]. The 10 databases included 316 lymphoma patients (targets) and 4692 non-lymphoma patients (controls). According to the data quality diagnostics (Fig. 2), a large portion of anomalies was detected in demographic- and laboratory-related measures, on each dataset, which were marked with orange color and removed from the analysis.

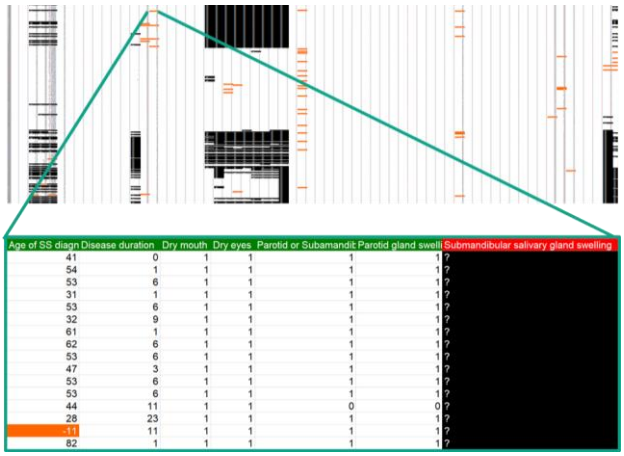

Figure 2. An instance of a selected dataset with quality diagnostics.

All features were ranked based on their quality. Instances with green color have adequate quality whereas those with red color have poor quality and fields with black color denote missing values (Fig. 2). A small portion (5%) of features with joint variability was identified between biopsy-related features. The flexible data harmonization approach yielded 41 features with more than 80% overlap across the 10 datasets.

B. Data augmentation

The density forest ensembles were applied on each dataset to augment the real population yielding 10,016 high-quality virtual patients (586 targets, 9430 controls), in total, with average gof 0.01, KL divergence less than 0.001, and correlation difference 0.02. The distributed learning pipeline was then utilized, using the hybrid loss function (Fig. 3), where the steepness of the logcosh and the wideness of the modified Huber loss were combined for different values. The value was defined as in the proposed distributed MART with dropouts, where r is the dropout rate.

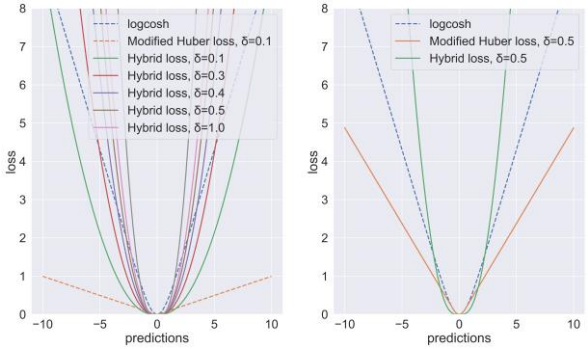

Figure 3. Distribution of the hybrid loss function compared to the modified Huber loss and the logcosh for different δ values.

In the real case, the 9 datasets having the highest number of targets were used for distributed training (300 targets and 4,411 controls, in total), whereas in the data augmentation case, the 9 training datasets included 9,422 patients (546 targets, 8,876 controls), in total. In both cases, the remaining (real) dataset was used for testing (16 targets, 281 controls). Random down-sampling with replacement was also applied on each case for class imbalance handling.

The overall performance of the distributed algorithms was better on the augmented data, where the distributed MART achieved accuracy 0.852, sensitivity 0.833 and specificity 0.854 against the one trained on the real data with accuracy 0.808, sensitivity 0.722 and specificity 0.818. A notable increase was observed in the case of the proposed distributed MART with $\delta = 0.4$ ($r = 0.3$) which achieved accuracy 0.865, sensitivity 0.84, and specificity 0.868 whereas in the real case the algorithm achieved accuracy 0.791, sensitivity 0.772, and specificity 0.794. A similar increase occurs for $\delta = 0.6$ ($r = 0.4$) with accuracy 0.862, sensitivity 0.868, and specificity 0.861 against the real case where the algorithm achieved accuracy 0.835, sensitivity 0.854, and specificity 0.833.

According to Fig. 4, the area under the curve scores in the distributed MART yielded an average increase by 5.2%, as well as, by 1.4% in the proposed distributed MART with $\delta = 0.4$ and 2.1% with $\delta = 0.6$.

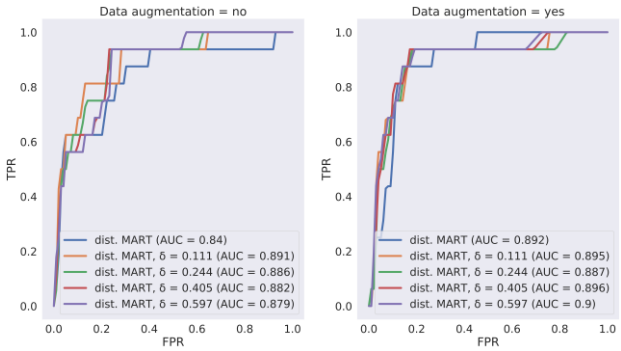

Figure 4. Receiver Operating Characteristic (ROC) curves for distributed classification with and without augmentation.

The positive impact of data augmentation is also reflected by the detection error tradeoff (DET) curves which are

depicted in Fig. 5 in logarithmic scale. The DET score was defined as the median absolute ratio of the false positive rate over the false negative rate. According to Fig. 5, an average decrease by 2.6% in the DET score is observed in the proposed distributed MART with $\delta = 0.4$ ($r = 0.3$) and 4.5% with $\delta = 0.6$ ($r = 0.4$).

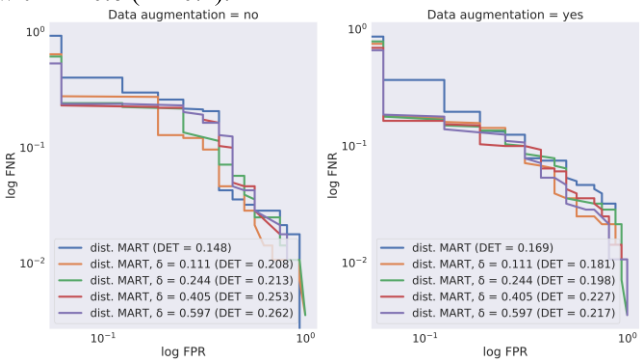

Figure 5. Detection error tradeoff (DET) curves for distributed classification with and without augmentation.

IV. Conclusions

In this work, we presented a pipeline for additive training across augmented and harmonized clinical data in distributed environments through the utilization of distributed multiple additive regression trees (MART) with a hybrid loss. The pipeline includes data pre-processing routines for the precise detection of data anomalies, as well as, features with joint variability. Both flexible and stringent lexical analysis were applied to detect terminologies with increased coherence among the distributed data. Density forest ensembles were finally developed for the generation of high-quality virtual distributions which were used for data augmentation.

The density forest ensembles were able to generate virtual data for data augmentation with decreased divergence with the real data (average gof 0.01, KL divergence less than 0.001, and correlation difference 0.02). The proposed pipeline was able to yield robust distributed learning models from the augmented data with an average increase by 6.8% in sensitivity, and 10.4% in specificity for $\delta = 0.4$. The proposed loss function avoids overfitting effects which are caused by the early inclusion of regression trees in the ensemble. To our knowledge, this is the first case study which combines data augmentation and distributed regression tree ensembles with hybrid loss yielding robust disease classification models through a case study in autoimmune diseases.

Although the proposed classifiers do not exist in distributed libraries like Apache Spark's MLlib [24] we plan to conduct a comparison study in the future. Moreover, we plan to extend the explainability of the classifiers by measuring the impact of each ensemble in the decision-making process and explore new utilities to avoid biases during the training stage.

References

[1] S. Shilo, H. Rossman, and E. Segal, "Axes of a revolution: challenges and promises of big data in healthcare," Nature medicine, vol. 26, no. 1, pp. 29-38, Jan. 2020.

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

[3] P. T. Chen, C. L. Lin, and W. N. Wu, "Big data management in healthcare: Adoption challenges and implications," International Journal of Information Management, vol. 53, pp. 102078, Jan. 2020.

[4] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, "Big data in healthcare: management, analysis and future prospects," Journal of Big Data, vol. 6, no. 1, pp. 1-25, June 2019.

[5] M. Tzanou, (Ed.), "Health Data Privacy Under the GDPR: Big Data Challenges and Regulatory Responses," Routledge, Nov. 2020.

[6] Y. Kumar, K. Sood, S. Kaul, and R. Vasuja, "Big data analytics and its benefits in healthcare," In: Kulkarni A. et al. (eds) Big Data Analytics in Healthcare, Springer, vol 66., Oct. 2019.

[7] A. Shastri and M. Deshpande, "A review of big data and its applications in healthcare and public sector," In: Kulkarni A. et al. (eds) Big Data Analytics in Healthcare, Springer, vol. 66, Oct. 2019.

[8] G. Lan, "First-order and Stochastic Optimization Methods for Machine Learning," Springer International Publishing, May 2020.

[9] W. U. Bajwa, V. Cevher, D. Papailiopoulos, and A. Scaglione, "Machine learning from distributed streaming data," IEEE Signal Process. Mag., vol. 37, no 3, pp. 11-13, May 2020.

[10] H. Li, S. Tian, Y. Li, Q. Fang, R. Tan, Y. Pan, C. Huang, Y. Xu, and X. Gao, "Modern deep learning in bioinformatics," Journal of molecular cell biology, pp. mjaa030, June 2020.

[11] D. P. Bertsekas, "Incremental gradient, subgradient, and proximal methods for convex optimization: A survey," Optimization for Machine Learning, vol. 2010, no. 1-38, pp. 3, Sep. 2011.

[12] 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 the European Medical and Biological Engineering Conference (EMBEC), vol. 80, pp. 306-313, Nov. 2020.

[13] V. C. Pezoulas, T. P. Exarchos, K. D. Kourou, A. G. Tzioufas, S. De Vita, and D. I. Fotiadis, "Utilizing incremental learning for the prediction of disease outcomes across distributed clinical data: A framework and a case study," In Proceedings of the Mediterranean Conference on Medical and Biological Engineering and Computing (MEDICON), pp. 823-831, Sep. 2019.

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

[15] 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.

[16] 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.

[17] Athena: OHDSI. URL: http://athena.ohdsi.org [accessed 2021-02-09].

[18] K. D. Kourou, V. C. Pezoulas, E. I. Georga, T. P. Exarchos, P. Tsanakas, M. Tsiknakis, T. Varvarigou, S. De Vita, A. G. Tzioufas, and D. I. Fotiadis, "Cohort harmonization and integrative analysis from a biomedical engineering perspective," IEEE reviews in biomedical engineering, vol. 12, pp. 303-318, Jul. 2018.

[19] Q. Wang, Y. Ma, K. Zhao, and Y. Tian, "A comprehensive survey of loss functions in machine learning," Annals of Data Science, pp. 1-26, Apr. 2020.

[20] Q. Sun, W. X. Zhou, and J. Fan, "Adaptive huber regression," Journal of the American Statistical Association, vol. 115, no. 529, pp. 254-265, Apr. 2019.

[21] 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.

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

[23] A. V. Goules, O. D. Argyropoulou, V. C. Pezoulas, L. Chatzis, E. Critselis, S. Gandolfo, et al., "Primary Sjögren's Syndrome of Early and Late Onset: Distinct Clinical Phenotypes and Lymphoma Development," Frontiers in immunology, vol. 11, pp. 2707, Oct. 2020.

[24] E. Nazari, M. H. Shahriari, and H. Tabesh, "BigData analysis in healthcare: apache hadoop, apache spark and apache flink," Frontiers in Health Informatics, vol. 8, no. 1, pp. 14, Jan. 2019.