Received 6 November 2019; revised 20 December 2019; accepted 20 December 2019. Date of publication 9 January 2020; date of current version 14 February 2020. The review of this paper was arranged by Editor Paolo Bonato.

Digital Object Identifier 10.1109/OJEMB.2020.2965191

Predicting Lymphoma Development by Exploiting Genetic Variants and Clinical Findings in a Machine Learning-Based Methodology With Ensemble Classifiers in a Cohort of Sjögren's Syndrome Patients

KONSTANTINA D. KOUROU [1,2], VASILEIOS C. PEZOULAS [1], ELENI I. GEORGA [1], THEMIS EXARCHOS[1,3], COSTAS PAPALOUKAS[1,2], MICHALIS VOULGARELIS[5], ANDREAS GOULES[5], ANDRIANOS NEZOS[6], ATHANASIOS G. TZIOUFAS[5], HARALAMPOS M. MOUTSOPOULOS[7], CLIO MAVRAGANI[5,6], AND DIMITRIOS I. FOTIADIS [1,4] (Member, IEEE)

[1]Unit of Medical Technology and Intelligent Information Systems, Department of Materials Science and Engineering, The University of Ioannina, GR45110 Ioannina, Greece
[2]Department of Biological Applications and Technology, The University of Ioannina, GR45110 Ioannina, Greece
[3]Department of Informatics, Ionian University, GR49100 Corfu, Greece
[4]Foundation for Research and Technology-Hellas, Institute of Molecular Biology and Biotechnology, Department of Biomedical Research, Ioannina GR45110, Greece
[5]Department of Pathophysiology, School of Medicine, National and Kapodistrian University of Athens, GR15772 Athens, Greece
[6]Department of Physiology, School of Medicine, National and Kapodistrian University of Athens, GR15772 Athens, Greece
[7]Academy of Athens, GR10679 Athens, Greece

CORRESPONDING AUTHOR: D. I. FOTIADIS (e-mail: fotiadis@cc.uoi.gr)

This work was supported in part by the European Union's Horizon 2020 research and innovation programme under Grant 731944 and in part by the Swiss State Secretariat for Education, Research and Innovation SERI under Grant 16.0210.
This article has supplementary downloadable material available at https://ieeexplore.ieee.org, provided by the authors.

ABSTRACT Lymphoma development constitutes one of the most serious clinico-pathological manifestations of patients with Sjögren's Syndrome (SS). Over the last decades the risk for lymphomagenesis in SS patients has been studied aiming to identify novel biomarkers and risk factors predicting lymphoma development in this patient population. *Objective:* The current study aims to explore whether genetic susceptibility profiles of SS patients along with known clinical, serological and histological risk factors enhance the accuracy of predicting lymphoma development in this patient population. *Methods:* The potential predicting role of both genetic variants, clinical and laboratory risk factors were investigated through a Machine Learning-based (ML) framework which encapsulates ensemble classifiers. *Results*: Ensemble methods empower the classification accuracy with approaches which are sensitive to minor perturbations in the training phase. The evaluation of the proposed methodology based on a 10-fold stratified cross validation procedure yielded considerable results in terms of balanced accuracy (GB: 0.7780 ± 0.1514, RF Gini: 0.7626 ± 0.1787, RF Entropy: 0.7590 ± 0.1837). *Conclusions:* The initial clinical, serological, histological and genetic findings at an early diagnosis have been exploited in an attempt to establish predictive tools in clinical practice and further enhance our understanding towards lymphoma development in SS.

INDEX TERMS Ensemble methods, genetic variants, lymphoma prediction, machine learning, Sjögren's Syndrome.

IMPACT STATEMENT We highlight the potential usefulness of genetic variants and clinical findings in predicting lymphoma development in Sjögren's Syndrome patients based on ensemble methods.

I. INTRODUCTION

Sjögren's syndrome (SS) is a chronic autoimmune disorder mainly manifested with dryness of mucosae as a result of exocrine gland involvement chiefly the salivary and lachrymal glands resulting in dry eyes and mouth. Systemic features also occur, in one third of the patients, as a result of skin, lungs, kidneys, liver and vessel involvement. Lymphoma development is one of the most serious manifestations. [1], [2].

Over the last decades a large amount of data revealed several clinical (salivary gland enlargement, purpura, Raynaud [3], [4]), hematological, serological (RF, Ro/La autoantibodies, monoclonal gammopathy [4]–[6], low complement C4 [3], serum BAFF [7], [8], sFLT [9]) and histopathological features (extensive lymphocytic infiltration [10]), as predictors for lymphoma development in Sjögren's syndrome. Of interest, these risk factors usually present at disease onset implying that a distinct genetic background could characterize the subgroup of SS patients which will develop lymphoma in the course of their disease [11].

On this basis, genetic variants of genes implicated in the regulation of chronic inflammation such as TNFAIP3 [12]–[14] and LILRA3 [15], B cell activation [16], [17], type I IFN pathways such as TREX-1 [18] as well as epigenetic processes [19] have been shown to increase the risk of Non-Hodgkin Lymphoma (NHL) in SS. The susceptibility to lymphoma development increases especially in patients in whom the disease starts before the age of 40 years old, as evidenced by the higher frequencies of the BAFF-R [17], [20], TNFAIP3 [12] and LILRA3 [15] variants.

Lymphoma prediction based on clinical and biological predictors have been studied in terms of statistical analysis and prediction rules [4], [21]–[24]. Towards this direction, in [4] a predictive tool in clinical practice has been developed for SS-related lymphoma development based on the initial clinical, laboratory and histopathological variables of SS patients Data mining algorithms have been also exploited for the identification of patient subgroups and the prediction of lymphoma in primary SS [25]. The associations among patient's demographics, clinical and serological variables have been defined and a prediction model based on Artificial Neural Networks (ANNs) has been developed able to predict new unseen records with high sensitivity and specificity. In the present study, we aim to identify the contribution of combined initial clinical, serological and histopathological features with genetic variants in predicting lymphoma development using Machine Learning-based (ML) methodology with ensemble classifiers. We focused on the development of a ML-based methodology able to classify accurately new patients according not only to their traditional clinical findings but also to their genetic susceptibility as a critical factor that predispose to lymphoma development in SS patients. The proposed methodology is based on the Gradient boosting (GB) [26] and Random Forest (RF) [27] ensemble classifiers for developing the predictive models which are characterized by the ability to generalize their decision boundaries to regions where there are no available training examples. This type of classifiers was selected in terms of the variance and bias estimation which contribute to the expected error of a classification model. The novelty of the proposed ML-based methodology pertaining to the potential usefulness of genetics in predicting lymphoma development in SS patients. The classification results reported in our study are obtained from stratified 10-fold cross validation with the ensemble classifiers outperforming the single Logistic Regression (LR) approach and the Support Vector Machine (SVM) classifier. Based on our results, we anticipate that the current work could provide new insights into the aggressive behavior of lymphoma development in SS patients.

II. RESULTS

Fig. 1 presents the evaluation performance of the GB and RF ensemble classifiers. More details are provided within the supplementary material regarding the obtained results, the study cohort, the preprocessing steps and the model training and parameter tuning of the ensemble classifiers. For the RF classifier both Gini and entropy criterion were applied in order to determine the best way to split the samples. These measures are defined according to the fraction of samples that belong to class i at a given node t. The best split is then selected according to the degree of impurity of the child nodes [32]. Three input cases were considered in the current study for comparison reasons and for assessing the models' performances (Table III supplementary material). More specifically, the clinical phenotype of each patient along with the genetic data were considered (input case 1) for building the proposed predictive models and further evaluate their performance. For assessing the potential of combining the initial SS patient's medical features with genetic variants in predicting lymphoma development, we followed the same procedure for input case 2 (the clinical phenotype for each patient) and input case 3 (the genotyped data acquired for each patient) and evaluated the models' performances in terms of certain metrics and hyperparameter optimization criterion (i.e., balanced accuracy). For each prediction model (i.e., RF models and GB model) the mean value of each metric is presented along with the computed standard deviation (Table III supplementary material).

We can observe that the combination of the initial clinical, serological and histopathological features with genetic variants result in the accurate prediction of lymphoma development in SS patients with considerable high balanced accuracy for RF Gini (0.7626 ± 0.1787), RF Entropy (0.7590 ± 0.1837) and GB (0.7780 ± 0.1514) classifiers, respectively (Table III supplementary material). We should also report for input case 1 (clinical and genetic data), the remarkable results obtained with reference to the sensitivity metric implying the high proportion of patients with lymphoma who have been predicted as positive by the classifiers (RF Gini classifier: 0.8000 ± 0.3435, RF Entropy classifier: 0.8000 ± 0.3435 and GB classifier: 0.8309 ± 0.2594) (Table III supplementary material and Fig. 1). As illustrated in the confusion matrices (Fig. 1), the GB model could predict more subjects as true positives (104) and true negatives (53) in comparison to RF Gini and RF models.

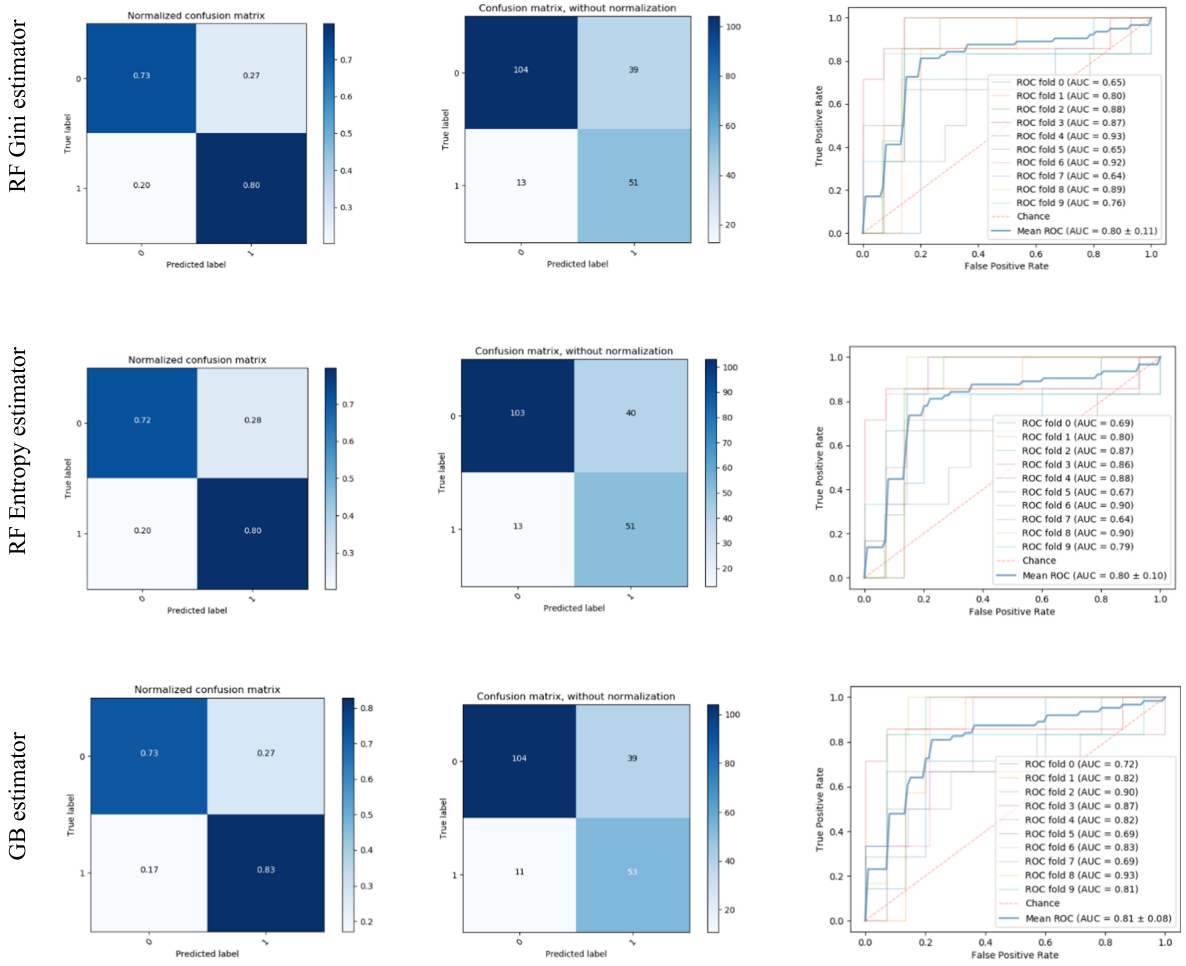

FIGURE 1. **The normalized and non-normalized confusion matrices obtained for each classification model. The ROC curves after the evaluation of models' performance are also illustrated. Each row corresponds to the respective classifier's evaluated performance. In the upper side the classification performance of RF Gini estimator is depicted (confusion matrices and ROC curve). In the middle and lower side of the figure the classification results of RF Entropy and GB classifiers are presented, respectively. The ROC curves correspond to the mean ROC curves and auc after applying the 10-fold cross validation procedure in the proposed ML methodology. The ROC curve in each fold is also illustrated for comparison purposes. In addition, the ± 1SD is also given with the mean ROC.**

The mean AUC of the models in terms of the sensitivity and specificity results are 0.7988 ± 0.2186 (RF Gini classifier), 0.7995 ± 0.1917 (RF Entropy classifier) and 0.8054 ± 0.1570 (GB classifier) which constitute promising results for predicting lymphoma development (Fig. 1).

For input case 2 (clinical data) the GB classifier performed better with slightly higher mean AUC (0.8215 ± 0.1534) in comparison to the mean AUC of input case 1 (clinical and genetic data) (0.8054 ± 0.1570). The exploitation of only the clinical patient records could be comparable with the combination of both genotyped data and the clinical phenotypes towards predicting lymphoma development. However, we can observe that the computed sensitivity, positive predictive and negative predictive values of the GB model for input case 2 (clinical data) are lower enough in accordance to the respective evaluation metrics for input case 1 (clinical and genetic data). Concerning the exploitation of individual genetic variants for building the predictive models (input case 3) the results yielded by the proposed methodology are moderate with significantly lower balanced accuracy, sensitivity and specificity in comparison to input case 1 and input case 2. Based on this knowledge, we can admit that the combination of both data sources (clinical and genetic profile) could result in more accurate classification results by obtaining predictive models with reference to ML techniques.

Fig. 2 illustrates the boxplot with mean feature importances according to the feature selection and ranking procedure performed with RF selector (section V. MATERIALS AND METHODS). Hence, the most important features which contribute to accurate and unbiased predictions of lymphoma development were identified. We can observe that the 10 most informative features are SGE, age at SS diagnosis, low C4, lymphadenopathy, RF plus, BAFF, TREX and MTHFR677 SNPs. We can observe that beyond these features, the most informative ones are mainly clinical findings and the rs11797 and rs12583006 reference numbers of the corresponding

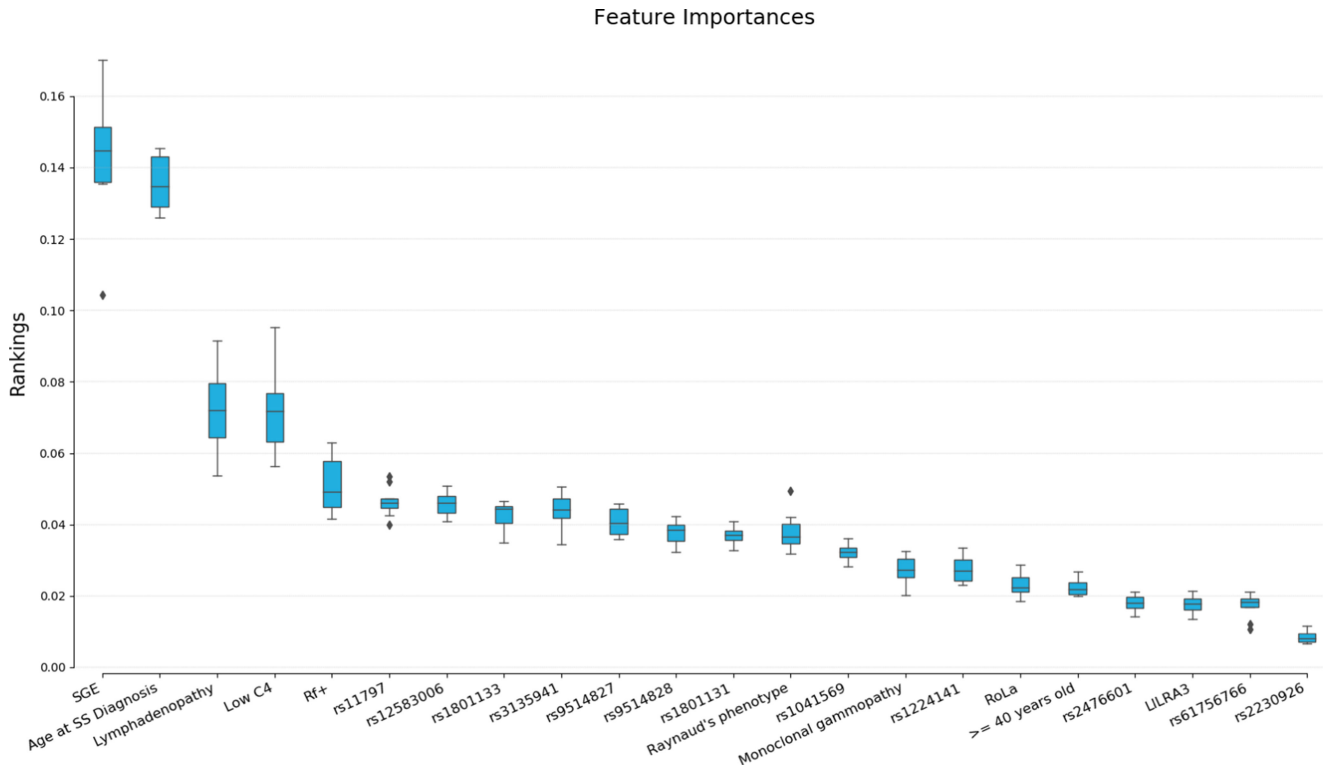

FIGURE 2. Boxplot with the mean feature rankings for each variable considered by the respective estimator. RF feature selection was performed with threshold the "mean" and "max_features" equal to the max number of features in the dataset considered at each experiment (input case 1clinical and genetic data).

genetic variants. We shall recall that the presented values refer to the mean importance rankings.

III. DISCUSSION

Predicting the risk for lymphoma development still remains a clinical unmet need in SS. The main clinical and genetic aspects of this major complication need to be elucidated for providing a meaningful clinical impact and translational findings in the field.

In this study, we highlight the potential of combining the clinical, serological and histological parameters along with the genetic profile of SS patients for the prediction of lymphoma development through a ML methodology consisting of ensemble algorithms. GB and RF classifiers were utilized to obtain accurate classification results based on their generalization ability and the minimization of errors in the training phase [33]–[36]. Based on the selected estimators in the inner ensemble, the training phase was conducted on different balanced bootstrap samples while random under-sampling was considered [34], [35]. Feature ranking was applied in terms of the RF selector based on importance weights. The threshold value used for feature selection and ranking was set to the maximum number of variables within our dataset. The number of features ranked by the estimator was 22, with SGE and age at SS diagnosis being the most important features that contribute to the classification of patients' samples (mean ranking of SGE = 0.1446, mean ranking of age at SS diagnosis = 0.1347). rs12583006 and rs11797 genetic variants are also included within the first 10 most informative features contributing to the prediction of lymphoma development (mean ranking of rs12583006 = 0.0462, mean ranking of rs11797 = 0.0460). The feature ranking results (Fig. 2) confirmed the identification of SGE and lymphadenopathy as independent adverse predictors for NHL development. We should also note that the age of patients at disease diagnosis could be a potential predictor for lymphoma development. According to published results, mucosa-associated lymphoid tissue (MALT) lymphoma occurs in younger pSS patients [37] which indicates the severity of diagnosis at an early stage. Furthermore, the rs11797 and rs12583006 genetic variants have been found as significant predictors along with specific clinical findings.

The mean ROC curves of RF Gini, RF Entropy and GB predictive models, with reference to input case 1 (clinical and genetic data), are also depicted in Fig. 3(a-c), including the variance of each curve based on the different subsets created when the training sets are splitted. The figures exhibit how the classifiers output is affected by changes in the training data and how different the subsets are from one another according to the cross-validation procedure. We can observe the low variance which is closely related to the robustness of our methodology. We can also observe the remarkably high results achieved by the three classifiers

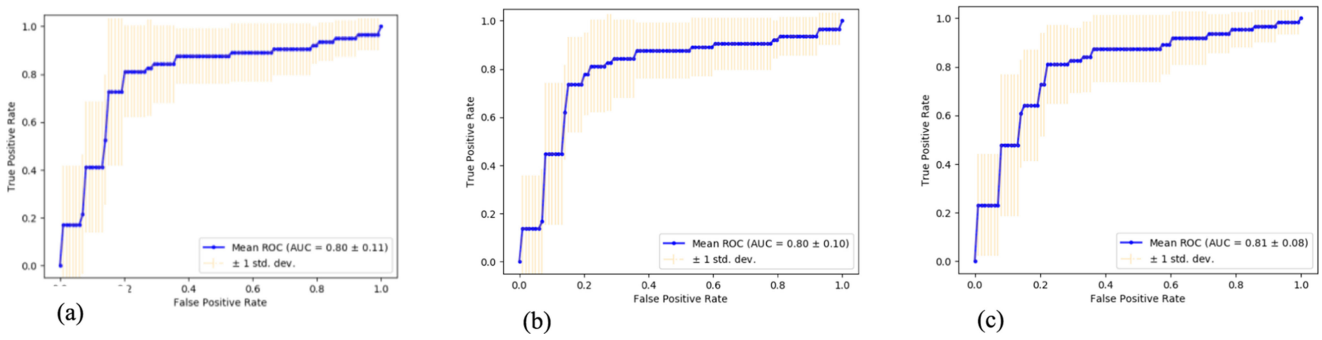

FIGURE 3. The calculated mean ROC curve and auc (a–c), with the variance of each curve when the training set is split into 10 different subsets. This pinpoints how the estimator output is affected by changes in the training data, and how different the splits are from one another in 10-fold cross validation. The left ROC curve corresponds to RF Gini estimator and the middle and right ones to RG Entropy and GB classifiers, respectively.

TABLE 1. The Variables of the Initial Demographic, Clinical and Laboratory Findings Related to the Patients' Samples Considered in the Current Study. The Mean±SD Values and the Min/Max Values Were Calculated for Continuous Variables. The Respective Percentages Were Also Calculated for the Discrete Variables. These Values Were Computed for Both Classes (I.E. Class 0 = No Lymphoma Development; Class 1 = Lymphoma Development). The Undefined Percentages for Categorical Variables are Also Given

Category	Variable	Class 0			Class 1		
		mean	SD	min / max	mean	SD	min / max
Demographic	Age at SS diagnosis (years)	50.93	13.38	15 / 74	50.67	14.27	24 / 81
Category	**Variable**	**Class 0 (%)**			**Class 1 (%)**		
		False	True	Undefined	False	True	Undefined
Clinical features	Salivary Grand Enlargement (SGE)	78.32	20.97	0.71	31.25	67.18	1.57
	Raynaud's phenomenon	78.35	21.65	0.00	59.37	40.63	0.00
	Lymphadenopathy	86.00	14.00	0.00	54.68	45.32	0.00
	≥ 40 age at SS diagnosis	82.52	17.48	0.00	25.00	75.00	0.00
Laboratory characteristics	Monoclonal gammopathy	89.51	6.29	4.20	71.87	25.00	3.13
	Anti-Ro/SSA or/and anti-La/SSB positivity	74.12	24.47	1.41	11.00	89.00	0.00
	RF positivity	50.34	41.25	8.41	15.62	82.81	1.57
	Low C4	53.14	45.45	1.41	20.31	76.56	3.13

in terms of the negative predictive value metric (Table III supplementary material). This constitutes a promising impact of our methodology in predicting accurately the patients that are found as negatives and actually do not have diagnosed with lymphoma during SS progression. As illustrated in Fig. 1, the high sensitivity values were obtained when both initial findings and genetic variants are exploited. This reveals the ability of the developed classification models to predict with high proportion the patients who have lymphoma and are truly predicted as positive. To evaluate the predictions on the test sets, different scores were also applied besides the balanced accuracy criterion, such as the f1 score, the log loss metric and the recall. However, the results obtained were similar or with very slight differences in comparison to the balanced accuracy scoring parameter. The proposed methodology was also applied to different input cases where the clinical and genetic variants were considered separately (Table III supplementary material). Obviously, the exploitation of the genotyped data from the patients result in moderate classification balanced accuracy related to the risk for lymphoma development.

On the contrary, individual clinical, serological and histopathological parameters have been identified in the literature as major predictors of B cell lymphomas. This is in accordance with the reported ML-based classification results (input case 2 in Table III supplementary material) revealing the superiority of collecting both the initial parameters and the genetic data on the disease onset. In the present work, we highlight the need for identifying risk clinical phenotypes in combination with the patients' genetic profiles for predicting the development of lymphomas which constitutes a major complication of SS. We show that the integration of both the patient's genetic background and the clinical phenotype could enhance the prediction accuracy of our ML models while improving disease diagnosis (Table III supplementary material). We further validated the methodology with other supervised learning methods used for classification, such as SVM (with linear kernel) and LR [4], [8]. Given the reported results based on the exploitation of both data types, we demonstrated that the proposed methodology with the ensemble classifiers outperforms the model performance based on SVM and LR. The reported balanced accuracy and AUC for SVM are 0.6395

TABLE 2. The Genetic Variants (Gene Ids and Rs Reference Numbers) Related to the Patients' Samples Considered in the Current Study. The Percentages for Common Genotype (0), Heterozygous (1), Homozygous (2) and Undefined SNPS Within Both Classes are Presented

Category	GENE/ ID	rs #	Class 0 (%)				Class 1 (%)			
			0	1	2	Undefined	0	1	2	Undefined
Genetic variant	MTHFR/ 4524	rs1801133	39.80	46.20	14.00	0.00	39.06	42.18	18.76	0.00
	MTHFR/ 4524	rs1801131	55.64	32.16	11.20	0.00	53.12	35.93	10.95	0.00
	TNFRSF13C (BAFF Receptor)/ 115650	rs61756766	93.70	6.30	0.00	0.00	90.60	9.40	0.00	0.00
	TNFSF13B (BAFF)/ 10673	rs1224141	71.32	26.57	1.40	0.71	60.93	35.93	0.00	3.14
	TNFSF13B (BAFF)/ 10673	rs12583006	51.04	35.66	13.30	0.00	53.12	39.06	4.70	3.12
	TNFSF13B (BAFF)/ 10673	rs9514828	15.39	56.64	27.97	0.00	10.93	56.25	29.68	3.14
	TNFSF13B (BAFF)/ 10673	rs1041569	52.44	44.05	3.51	0.00	51.56	39.06	6.25	3.13
	TNFSF13B (BAFF)/ 10673	rs9514827	44.05	44.75	11.20	0.00	35.93	48.43	10.93	4.71
	TREX1/ 11277	rs11797	35.66	44.75	18.90	0.69	34.37	40.62	23.43	1.58
	TREX1/ 11277	rs3135941	69.23	23.07	7.00	0.70	78.12	17.18	3.12	1.58
	TNFAIP3/ 7128	rs2230926	89.51	9.80	0.00	0.69	93.75	6.25	0.00	0.00
	PTPN22/ 26191	rs2476601	89.51	10.49	0.00	0.00	89.06	10.94	0.00	0.00
	LILRA3/11026	deletion	79.00	14.70	0.00	6.30	87.50	6.25	1.56	4.69

$\pm$ 0.2540 and 0.6934 $\pm$ 0.2586, respectively. The evaluated performance for the LR predictive model resulted in balanced accuracy 0.7259 $\pm$ 0.2087 and AUC 0.7962 $\pm$ 0.2133.

Based on the scientific studies published in the field which deal with the underlying factors and mechanisms that predispose lymphoma occurrence [9]–[14], we could state that the proposed work constitutes a complementary work with considerable prediction results. Although novel biomarkers have been identified (i.e., BAFF and TNFAIP3 polymorphisms) and validated risk scores have been also developed in terms of clinical parameters [4], [12], [13], [15], we showed that the combination of both data types and the application of ML-based frameworks could result in robust predictive models with impact in the clinical practice. In the era of precision medicine, the exploitation of heterogeneous data types could reveal new knowledge related to the complex molecular mechanisms of cancer development. The relatively small number of SS patients and the class imbalanced problem related to class 1 (i.e., 64 with either a history or a current diagnosis of SS NHL) are the main limitations of the current study. However, given the rates of unrecognized diagnosis of SS patients in the general population as well as the infrequency of SS initial findings in the healthcare sector, the dataset of the present study can be considered as one of the largest SS databases.

IV. MATERIALS AND METHODS

A. DATA COLLECTION AND CURATION

1) STUDY COHORT

Medical records of 143 primary SS patients (SS) without and 64 SS patients with a history or a current diagnosis of B-cell Non-Hodgkin lymphoma (SS NHL), fulfilling the revised European/American International classification criteria for SS, were collected (Table 1 and supplementary material). DNA derived from whole peripheral blood of 207 patients with primary SS fulfilling the revised European/American classification criteria [28] was collected. The patients were genotyped for 13 single nucleotide polymorphisms (Table 2) after was extracted and stored at -20°C upon use at the Department of Physiology, National and Kapodistrian University of Athens, Athens, Greece. Methods of DNA extraction and genotyping protocols have been previously described in [12], [15]–[19].

Demographic, clinical and laboratory features were recorded after thorough chart review. Lymphoma diagnosis in the SS-lymphoma group was based on the criteria outlined by the World Health Organization classification. This study was carried out in accordance with the recommendations of the Ethics Committee of the National and Kapodistrian University of Athens (approved No. 6337) with written informed consent from all subjects following the Declaration of Helsinki.

2) DATA PREPROCESSING AND CURATION

Data preprocessing was performed by utilizing an automated framework for evaluating the data quality [29]. The main steps followed towards the dataset quality assessment are referred to the detection of (i) missing values in an autonomous way, (ii) removal of outliers, and (iii) duplicate values and highly correlated distributions among variables. More details are provided within the supplementary material with reference to the preprocessing steps and the data curation procedure that were followed.

B. COST-SENSITIVE RANDOM FOREST FEATURE SELECTION AND RANKING

The RF classifier was applied aiming at evaluating the importance of features with reference to the classification problem (supplementary material). The "balanced mode" of the RF estimator was selected in the current study to automatically adjust weights associated with the class frequencies in the training set. The identification of the most important predictor variables which contribute to accurate and unbiased predictions of the response variable was achieved. The maximum number of features selected after keeping the threshold disabled (i.e., threshold $= -\infty$) was also reported with reference to the feature ranking results.

C. ENSEMBLE METHODS

Ensemble methods enhance the classification accuracy by aggregating the predictions of multiple base classifiers [32]. During a classification task with ensemble methods a set of base classifiers is developed from the training data and the performance of the classification model is evaluated by voting on the individual predictions made by each classifier. The rationale for ensemble methods is that the error rate during a classifier's performance is considerably lower than the error rate of the base classifiers, considering that the base classifiers are not identical but independent [32].

Let D denote the original training data and T be the test set. A training set D_i is created from D, which size is kept identical with the original data while the distribution of records may be different. A base learner C_i is built from D_i, for $i = 1, \ldots, k$, which denotes the number of base classifiers. For each test record $x \in T$ to be classified, the predictions made by each base classifier $C_i(x)$ are then aggregated by taking a majority vote on the individual base learners predictions in order to obtain the class $C^*(x)$:

$$C^*(x) = Vote\left(C_1(x),\ C_2(x),\ \ldots,\ C_k(x)\right) \qquad (1)$$

Ensemble methods achieve better classification results with unstable classifiers which are sensitive to minor perturbations in the training phase. Examples of such classifiers are the decision trees, the rule-based classifiers and the artificial neural networks [32]. The proposed ML-based methodology enables the minimization of errors related to the variability of the training samples due to the utilization of ensemble algorithms. The bias-variance decomposition method is usually applied for the analysis of such types of errors concerning the predictions of a classification model [32]. In the current study, the GB and RF ensemble classifiers are considered and further implemented [33]–[36] based on imbalanced datasets which consist of categorical variables. We aim to develop predictive models, in terms of machine learning techniques, with high generalization ability and less training errors. More details are given in the supplementary material related to the GB and RF classification models, the performance evaluation and validation along with their parameter tuning.

V. CONCLUSIONS

According to the reported classification results in the current study we could conjecture about the potential of exploiting the clinicogenomic profiles of patients for predicting lymphoma development during SS progression. Based on the proposed ML-based methodology we demonstrated that ensemble methods could obtain promising classification results comparing to conventional statistical methods and/or other supervised learning algorithms used for the development of predictive models in healthcare. Although lymphoma development presents an unmet clinical need in the research field of SS, the international efforts among groups and the conduction of SS prospective studies could provide a clinical impact to the disease management and the patients' daily activity. Apparently, Genome Wide Association studies (GWAs) could provide observational studies of genome-wide genetic variants which can be easily incorporated in our proposed methodology; thus, enhancing the identification of new population-based risk genetic variants in SS. Towards this direction, the exploitation of large and heterogeneous SS datasets in future multicenter studies could contribute to the development of more accurate predictive models through ML techniques. Furthermore, the rise of omics data and their exploitation in the biomedical sciences could empower the identification of key factors involved in lymphomagenesis and the detection of high-risk patients at early stages.

SUPPLEMENTARY MATERIALS

In the supplementary material details are given related to the obtained results, the study cohort, the preprocessing steps (i.e. data curation) and the model training and parameter tuning of the ensemble classifiers.

REFERENCES

[1] C. P. Mavragani and H. M. Moutsopoulos, "Sjögren syndrome," *Cmaj*, vol. 186, pp. E579–E586, 2014.

[2] E. Zintzaras, M. Voulgarelis, and H. M. Moutsopoulos, "The risk of lymphoma development in autoimmune diseases: A meta-analysis," *Archives Internal Med.*, vol. 165, pp. 2337–2344, 2005.

[3] F. N. Skopouli, U. Dafni, J. P. Ioannidis, and H. M. Moutsopoulos, "Clinical evolution, and morbidity and mortalityof primary Sjögren's syndrome," in *Proc. Seminars Arthritis Rheumatism*, 2000, pp. 296–304.

[4] 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, 2016, Art. no. e3766.

[5] G. Nocturne *et al.*, "Rheumatoid factor and disease activity are independent predictors of lymphoma in primary Sjögren's syndrome," *Arthritis Rheumatology*, vol. 68, pp. 977–985, 2016.

[6] A. L. Tomi *et al.*, "Brief report: Monoclonal gammopathy and risk of lymphoma and multiple myeloma in patients with primary Sjögren's syndrome," *Arthritis Rheumatology*, vol. 68, pp. 1245–1250, 2016.

[7] L. Quartuccio *et al.*, "BLyS upregulation in Sjögren's syndrome associated with lymphoproliferative disorders, higher ESSDAI score and B-cell clonal expansion in the salivary glands," *Rheumatology*, vol. 52, pp. 276–281, 2012.

[8] A. Nezos *et al.*, "Type I and II interferon signatures in Sjogren's syndrome pathogenesis: Contributions in distinct clinical phenotypes and Sjogren's related lymphomagenesis," *J. Autoimmunity*, vol. 63, pp. 47–58, 2015.

[9] G. J. Tobón *et al.*, "The Fms-like tyrosine kinase 3 ligand, a mediator of B cell survival, is also a marker of lymphoma in primary Sjögren's syndrome," *Arthritis Rheumatism*, vol. 62, pp. 3447–3456, 2010.

[10] A. P. Risselada, A. A. Kruize, R. Goldschmeding, F. P. Lafeber, J. W. Bijlsma, and J. A. van Roon, "The prognostic value of routinely performed minor salivary gland assessments in primary Sjögren's syndrome," *Ann. Rheumatic Diseases*, vol. 73, pp. 1537–1540, 2014.
[11] A. Nezos and C. P. Mavragani, "Contribution of genetic factors to Sjögren's syndrome and Sjögren's syndrome related lymphomagenesis," *J. Immunology Res.*, vol. 2015, 2015, Art. no. 754825.
[12] A. Nezos, E. Gkioka, M. Koutsilieris, M. Voulgarelis, A. G. Tzioufas, and C. P. Mavragani, "TNFAIP3 F127C coding variation in Greek primary Sjogren's syndrome patients," *J. Immunology Res.*, vol. 19, 2018, Art. no. 6923213.
[13] G. Nocturne *et al.*, "Germline and somatic genetic variations of TNFAIP3 in lymphoma complicating primary Sjögren's syndrome," *Blood*, vol. 122, pp. 4068–4076, 2013.
[14] G. Nocturne *et al.*, "Germline variation of TNFAIP3 in primary Sjögren's syndrome-associated lymphoma," *Ann. Rheumatic Diseases*, vol. 75, pp. 780–783, 2016.
[15] E. Argyriou *et al.*, *THU0204 Association of Lilra3 Gene With Lymphomagenesis Risk in Young SS Patients*, ed: Ann Rheum Dis. Scientific abstracts: BMJ Publishing Group Ltd, 2019.
[16] A. Nezos *et al.*, "B-cell activating factor genetic variants in lymphomagenesis associated with primary Sjogren's syndrome," *J. Autoimmunity*, vol. 51, pp. 89–98, 2014.
[17] A. Papageorgiou *et al.*, "A BAFF receptor His159Tyr mutation in Sjögren's Syndrome–related lymphoproliferation," *Arthritis Rheumatology*, vol. 67, pp. 2732–2741, 2015.
[18] A. Nezos *et al.*, "TREX1 variants in Sjogren's syndrome related lymphomagenesis," *Cytokine*, vol. 17, 2019, Art. no. 154781.
[19] S. Fragkioudaki *et al.*, "MTHFR gene variants and non-MALT lymphoma development in primary Sjogren's syndrome," *Sci. Rep.*, vol. 7, pp. 1–8, 2017.
[20] C. P. Mavragani, A. Nezos, I. Sagalovskiy, S. Seshan, K. A. Kirou, and M. K. Crow, "Defective regulation of L1 endogenous retroelements in primary Sjogren's syndrome and systemic lupus erythematosus: Role of methylating enzymes," *J. Autoimmunity*, vol. 88, pp. 75–82, 2018.
[21] A. V. Goules and A. G. Tzioufas, "Lymphomagenesis in Sjögren's syndrome: Predictive biomarkers towards precision medicine," *Autoimmunity Rev.*, vol. 18, 2018, pp. 137–143.
[22] G. Nocturne, E. Pontarini, M. Bombardieri, and X. Mariette, "Lymphomas complicating primary Sjögren's syndrome: From autoimmunity to lymphoma," *Rheumatology*, 2019, doi: 10.1093/rheumatology/kez052.
[23] L. Quartuccio *et al.*, "Biomarkers of lymphoma in Sjögren's syndrome and evaluation of the lymphoma risk in prelymphomatous conditions: Results of a multicenter study," *J. Autoimmunity*, vol. 51, pp. 75–80, 2014.
[24] J. P. Ioannidis, V. A.Vassiliou, and H. M. Moutsopoulos, "Long-term risk of mortality and lymphoproliferative disease and predictive classification of primary Sjögren's syndrome," *Arthritis Rheumatology*, vol. 46, no. 3, pp. 741–747, 2002.
[25] C. Baldini, F. Ferro, N. Luciano, S. Bombardieri, and E. Grossi, "Artificial neural networks help to identify disease subsets and to predict lymphoma in primary Sjögren's syndrome," *Clin. Exp. Rheumatology*, vol. 36, pp. 137–144, 2018.
[26] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001.
[27] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.
[28] I. E. Lundberg *et al.*, "2017 European League against rheumatism/American college of rheumatology classification criteria for adult and juvenile idiopathic inflammatory myopathies and their major subgroups," *Arthritis Rheumatology*, vol. 69, pp. 2271–2282, 2017.
[29] V. C. Pezoulas *et al.*, "Medical data quality assessment: On the development of an automated framework for medical data curation," *Comput. Biol. Med.*, vol. 107, pp. 270–283, 2019.
[30] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
[31] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts, "Understanding variable importances in forests of randomized trees," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 431–439.
[32] P.-N. Tan, *Introduction to Data Mining*. India: Pearson Education, 2018.
[33] J. H. Friedman, "Stochastic gradient boosting," *Comput. Statist. Data Anal.*, vol. 38, pp. 367–378, 2002.
[34] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, pp. 559–563, 2017.
[35] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Syst., Man, Cybern., B (Cybern.)*, vol. 39, pp. 539–550, 2008.
[36] C. Chen, A. Liaw, and L. Breiman, *Using Random Forest to Learn Imbalanced Data*," Berkeley: Univ. California, vol. 110, p. 4, 2004.
[37] A. Papageorgiou *et al.*, "Predicting the outcome of Sjogren's syndrome-associated non-Hodgkin's lymphoma patients," *PLoS One*, vol. 10, 2015, Art. no. e0116189.