Supervised machine learning techniques applied to medical records toward the diagnosis of rare autoimmune diseases

Pedro Emilio Andrade Martins[1, *] [0000-0002-4821-7968], Márcio Eloi Colombo Filho[1] [0000-0003-3779-0192], Ana Clara de Andrade Mioto[1] [0000-0003-4475-1984], Filipe Andrade Bernardi[1] [0000-0002-9597-5470], Vinícius Costa Lima[1] [0000-0002-2467-358X], Têmis Maria Félix[2] [0000-0002-8401-6821], and Domingos Alves[1,3] [0000-0002-0800-5872]

[1] Health Intelligence Laboratory, Ribeirão Preto Medical School, Ribeirão Preto, Brazil
[2] Medical Genetics Service, Porto Alegre Clinical Hospital, Porto Alegre, Brazil
[3] Department of Social Medicine, Ribeirão Preto Medical School, Ribeirão Preto, Brazil

* Corresponding author: pedroemilioam02@usp.br

Abstract. Rare autoimmune diseases provoke immune system malfunctioning, which reacts and damages the body's cells and tissues. They have a low prevalence, classified as complex and multifactorial, with a difficult diagnosis. In this sense, this work aims to support the diagnosis of a rare autoimmune disease using the analysis of medical records from supervised machine learning methods and to identify the models with the best performance considering the characteristics of the available data set. A synthetic database was created with 1000 samples from epidemiological studies identified in the literature, simulating demographic data and symptoms from patient records for the diagnosis of Amyotrophic Lateral Sclerosis (ALS), Multiple Sclerosis (MS), Systemic Lupus Erythematosus (SLE), Crohn's Disease (CD) and Autoimmune Hepatitis (AIH). Data were segmented into training (80%) and test (20%), assigning the diagnosis as the class to be predicted. The models with the highest accuracy were Logistic Regression (84%), Naive-Bayes (82.5%), and Support Vector Machine (79%). Only LR obtained quality values greater than 0.8 in all metrics. SLE had the highest precision and recall for all classifiers, while ALS had the worst results. The LR accuracy model had the best performance with good quality metrics, although it was impossible to predict ALS accurately. Although a real dataset was not used, this work presents a promising approach to support the challenging diagnosis of rare autoimmune diseases.

Keywords: Machine learning, Rare diseases, Medical records.

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_13

1 Introduction

The concept of autoimmune diseases is derived from the improper functioning of the immune system. The defense cells initiate an erroneous process of attacking the body's substances since it recognizes them as possible antigens. It initiates an inflammatory cascade effect and destruction of the target component. The lymphocyte mistakenly recognizes its molecule as an antigen, continuously stimulating an immune response and directly affecting the functioning of tissues and other body structures [1].

Although the pathophysiological procedure of autoimmune diseases and their mechanisms of action is widely recognized, the same causal factor that triggers autoimmune diseases is unknown. However, studies suggest that genetic aspects, hormonal dysregulation, and environmental factors combine in an interdependent way in the manifestation of these diseases, allowing the understanding of these as complex multifactorial [2].

According to the medical literature, it is estimated that about 3% to 5% of the world's population suffers from an autoimmune disease, which is more common in women [3]. Some already documented diseases are Type 1 Diabetes Mellitus, Psoriasis, Multiple Sclerosis, Rheumatoid Arthritis, Systemic Lupus Erythematosus, and others [4].

Diagnosing autoimmune diseases is often complex and difficult to be carried out assertively, given the reaction similarity between the diseases in their initial stages. The distinction between symptoms and laboratory tests is insufficient, requiring a comprehensive investigation of the patient's family health history, medical imaging exams, and the study of previous medical events and symptoms [5].

The medical experience is a predominant factor, directly influencing decision-making in carrying out specific tests and, consequently, in the diagnostic process. In this way, although there are key indicators to guide the possibilities during the investigation, the diagnosis is not fully supported by tests but also by what is most plausible and likely given the combination of the physician's intrinsic knowledge, literature, and information obtained from the patient and the exams.

According to the Brazilian Ministry of Health, rare diseases are categorized as those with a prevalence of up to 65 people per 100.000 inhabitants [6]. Rare autoimmune diseases fit into this scenario according to epidemiological studies in Brazil and other countries [6-10].

Rare autoimmune diseases accentuate the problem regarding its assertive diagnosis due to the similarity between symptoms, the need for different laboratory tests and medical images, and professional experience. Examples such as Amyotrophic Lateral Sclerosis (ALS), Multiple Sclerosis (MS), Systemic Lupus Erythematosus (SLE), Crohn's Disease (CD), and Autoimmune Hepatitis (AIH) accurately represent this complexity, as they denote a costly and time-consuming diagnosis, with a severe prognosis, besides their clinical and epidemiological significance in Brazil.

According to the literature, the most recurrent uses of machine learning and deep learning aimed at autoimmune diseases are for predicting disease progression and diagnosis [11]. Also, the identification of possible biomarkers for the detection and formulation of inhibitory drugs and the prediction of candidate genetic sequences to be interpreted as autoantigens is a growing field [12, 13].

It is reported a wide variation of data types applied to these methods with a main focus on clinical data, especially magnetic resonance images, and genomic data [11]. Few studies have addressed the diagnosis between multiple autoimmune diseases, and these utilized rRNA gene data. [14, 15].

No related studies were found regarding the correlation of the main topics as machine learning techniques towards the diagnosis of multiple rare autoimmune diseases and medical records as the main data source. Therefore, this research provides an early comprehension about the subject, with relevance on avoiding the diagnosis delay by possibly identifying features on medical records and guiding diagnosis confirmation tests.

2 Objective

This research aims to study the performance of machine learning models as a decision support tool in the diagnosis of a rare autoimmune disease, based on patient medical records

3 Methods

3.1 Database Elaboration

The database is an essential component to carry out analyses and inferences toward the objective of this research. A database was developed for the project, simulating data from the patient's medical records due to the lack of relevant public data related to rare autoimmune diseases.

First, it is well established that criteria are needed to build the data set to simulate information about patient records and rare autoimmune diseases more faithfully with reality. The data structure was based on studies and systematic reviews regarding the epidemiology of rare autoimmune disorders [16-24]. Thus, the following parameters were applied:

Epidemiological Studies. For the selection of reviews and epidemiological studies as a reference, the authors put the following reasoning into practice: Limit the search to only five rare autoimmune diseases, namely: ALS, MS, SLE, CD, and AIH, since they have clinical, epidemiological relevance in Brazil; Prioritize epidemiological studies on the Brazilian population; Search for systematic reviews of the global epidemiology.

Sample Set. As these are rare diseases, patient medical records data is highly scarce. Given this, to determine the number of samples (n) to create the database, a systematic review with meta-analysis of Autoimmune Hepatitis was used as a reference [25]. Therefore, the elaborated database has n = 1000.

Variables Selection. From the data that make up the patient's medical records according to the literature, together with those from the epidemiological studies surveyed, the selected attributes were: Patient Identification; Age; Race; Sex; Symptoms, and Diagnosis. The attributes and structure of the dataset are shown in Table 1.

Table 1. Dataset Attributes.

Attribute name	Meaning	Attribute Type
ID	Patient Identification	Discrete
Sex	Gender	Categorical
Race	Ethnicity	Categorical
Age	Age arranged in time interval	Categorical
Symptom_F	Fatigue Symptom	Boolean
Symptom_M	Stiffness and loss of muscle strength	Boolean
Symptom_W	Loss weight	Boolean
Symptom_GI	Gastrointestinal Symptom	Boolean
Symptom_VD	Visual disorder	Boolean
Symptom_SI	Skin injuries	Boolean
Diagnosis	Rare autoimmune disease diagnosis	Categorical

Values Parameters. The following epidemiological data were used to determine how the values within each attribute would be distributed:

Prevalence. Used to confirm the disease as a rare autoimmune disease, specify the frequency of each diagnosis within the database, and verify which rare diseases are more common or less common.

Prevalence by Sex. Used to indicate the frequency of female or male patients by disease (MS, ALS, SLE, CD, AIH).
Prevalence by Race. Used to show the distribution of races by disease.

Graphs of distribution of age groups to the detriment of diagnosis. Used to indicate the percentage of age groups in each disease, applying it to our sample set.

Symptoms. Used to map lists of symptoms for each disease and trace common ones.

Furthermore, for the symptoms' selection, inclusion criteria were established that symptoms must be present in 3 or more diseases, as shown in Table 2: be a frequent symptom; be a generalized symptom; and finally, a patient must present at least two symptoms that match the characteristics of a given disease.

Table 2. Distribution of symptoms.

Symptoms	MS	ALS	SLE	CD	AIH
Symptom_F	Present	Present	Present	Present	Present
Symptom_M	Present	Present	Present	Absent	Absent
Symptom_W	Present	Present	Present	Present	Present
Symptom_GI	Present	Present	Present	Present	Present
Symptom_VD	Present	Present	Present	Present	Absent
Symptom_SI	Absent	Absent	Present	Present	Present

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_13

Thus, from these specifications, whose objective is to guarantee the greatest possible basis for the creation of the database, data were created and randomly allocated between the attributes and the class.

Data Processing. With selecting the sample set, it is crucial to process the data to avoid bias and unbalanced values. However, since the authors authored the dataset and it was standardized, there were few steps necessary to process the data.

Among the data pre-processing activities, the “Age” attribute was changed from discrete values to age range intervals to assist in the analysis process as categorical data. The “ID” attribute was removed since it has no significance among the other attributes and may even interfere during the analysis and classification process of the algorithms.

Finally, the dataset was segmented into two groups: the training and test sets. This division was made at random, following the proportion of 80% for the training set and 20% for the test stage of the models. In this scenario, the diagnosis of the autoimmune disease present in the patient's medical record is the "Class" to be predicted by machine learning algorithms.

3.2 Analysis Software

The Weka 3 software, a data mining tool, was chosen to analyze data by applying supervised machine learning algorithms. This software was chosen because it presents classification, regression, clustering, data visualization techniques, and various classification algorithms. Also, Weka 3 has helpful quality metrics for validating the machine learning model, such as the PRC Area, Confusion Matrix, and the F-Score.

3.3 Classification Algorithms

Supervised learning algorithms were used to select classification algorithms, which are already described in the literature for presenting good results in data analysis from patient records and for data related to rare diseases, especially Support Vector Machine and Artificial Neural Networks [26]. KNN, Naive-Bayes, Random Forest and Logistic Regression, has previously been used on MS diagnosis and others neurodegenerative diseases [27, 28]

All algorithms were used with default parameters. No optimal hyperparameters were selected, since this study aims at the initial comparison between well described classifiers.

The algorithms to which the dataset will be applied are shown below.

Support Vector Machine (SVM). Supervised machine learning algorithm applied for classification and regression. It consists of building a set of hyperplanes in an 'n-dimensional space, which allows a better distinction between classes using the maximum values between each class for its delimitation [29].

KNN. The supervised learning algorithm, known as 'lazy,' is used for dataset classification and regression. A class separation method based on a 'K' parameter selects the closest variables in space by the point-to-point distance (e.g., Euclidean, Hamming,

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_13

Manhattan). Given the class of these close variables, the model classifies from the one that appears most in the set within a certain distance [30].

Naive-Bayes (NB). The supervised classifier uses Thomas Bayes' theorems to ensure a simplification approach to the problem, being expressively relevant in real-world scenarios in the health area. This model assumes that all attributes are independent and relevant to the result, generating probabilities and frequency of existing values in comparison with the class to be predicted [31].

Random Forest (RF). Supervised learning algorithm executed for classification and regression problems, being able to use data sets containing continuous and categorical variables. This model uses decision trees in different samples, predicting the classification and regression from most data or its mean [32].

Logistic Regression (LR). The statistical model is used in supervised machine learning, allowing classification. It combines the attributes present in the dataset with mathematical and statistical methods to predict a class. In addition, it allows for analyzing the relationships between the present attributes, such as autocorrelation and multicollinearity, verifying the variables that best explain the expected output [33].

Multilayer Perceptron (MLP). It is an artificial neural network with one or more hidden layers and an indeterminate number of neurons. It consists of non-linear mathematical functions based on the backpropagation technique, which trains neurons by changing synaptic weights at each iteration [34].

3.4 Quality Metrics

First, all models presented the same evaluation metrics as the tables in the results. The most significant measures to qualify the performance of the classifiers were: Precision, Recall, F-score, Matthews Correlation Coefficient (MCC), and PRC Area.

Precision and Recall are fundamental validation statistics, as they are based on the number of true positive (PV) ratings, in other words, how many individuals were correctly categorized according to the class. The F-score is relevant because it synthesizes the Precision and Recall rates from a harmonic mean.

As it deals with diagnosis and differentiation between classes of diseases, the MCC is a highly relevant metric since it only presents a satisfactory result if all categories of the confusion matrix obtain good values, highlighting cases of false negative (FN) and false positive (FP). Finally, the PRC Area allows the graphical analysis of the relationship between precision and sensitivity, being preferable in comparison to the ROC Area because it is an unbalanced data set.

4 Results

4.1 Training set

First, due to the segmentation of the dataset between the training and test groups, a sample set n = 800 (80%) was used for the training process of the classification

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36024-4_13

algorithms. The distribution between classes in the training set is shown in Table 3. The set was applied to each classifier using cross-validation with K folds = 10, obtaining the results presented in the following sections.

Table 3. Distribution of Classes in Training Set.

Class	Sample	Percentage
MS	117	14.6%
ALS	43	5.4%
SLE	321	40%
CD	206	25.8%
AIH	113	14.2%

4.2 Test set

The test set was randomly chosen, selecting the first n = 200 (20%) patients. The distribution between classes is shown in Table 4.

Table 4. Distribution of Classes in Test Set.

Class	Sample	Percentage
MS	33	16.5%
ALS	7	3.5%
SLE	79	39.5%
CD	44	22%
AIH	37	18.5%

The test set was applied to each of the previously described models without any parameter changes for validation and verification of the performance with the new data.

4.3 Applied Classifiers in the Test Set

Support Vector Machine (SVM). From inserting new data, the SVM model presented a correct classification of 158 patients (79%) and weighted mean precision equivalent to 0.811. Such values were similar to those obtained during training, 0.783 and 0.790, respectively. Furthermore, the model had an average recall between classes of 0.790 and an F-score of 0.795.

From a class perspective, there was a considerable variation in metrics. In comparison, the SLE PRC Area was 0.970. The others did not have a value greater than 0.650, with a lower value of 0.312 for ALS. As shown in Table 5, other metrics highlight the imbalance between classes.

Table 5. SVM Detailed Performance by Class.

TP Rate	Precision	Recall	F-Measure	MCC	PRC Area	Class

DOI: 10.1007/978-3-031-36024-4_13

0.899	1,000	0.899	0.947	0.918	0.970	SLE
0.571	0.667	0.571	0.615	0.604	0.312	ALS
0.773	0.618	0.773	0.687	0.592	0.577	DC
0.606	0.800	0.606	0.690	0.647	0.594	MS
0.784	0.674	0.784	0.725	0.660	0.638	AIH
0.790	0.811	0.790	0.795	0.743	0.737	Weighted Avarage

KNN. The KNN models with K = 3, K = 5, and K = 7 did not show disparity between the metrics, with minimal to no difference when detailed by Class. Therefore, as seen in Table 6, we only focused on the comparison between the overall performance of this models.

The three models demonstrated an accuracy of 73.5% (n = 147), with more excellent recall for KNN-7 and KNN-5 (0.735) compared to KNN-3 (0.730). The KNN-7 provided slightly better results than the other two models, with an F-score of 0.735 and an accuracy of 0.749.

Table 6. KNN Average Performance.

TP Rate	Precision	Recall	F-Measure	MCC	PRC Area	Classifier
0.730	0.742	0.730	0.730	0.660	0.793	KNN-3
0.735	0.744	0.735	0.733	0.666	0.811	KNN-5
0.735	0.749	0.735	0.735	0.668	0.812	KNN-7

Naive-Bayes (NB). NB correctly predicted 165 patients (82.5%) with a mean across-class accuracy of 0.836. It also revealed a MCC of 0.783 and an Area PRC of 0.861.

As shown in Table 7, the classes with the highest recall are, respectively, SLE (0.924), AIH (0.825) and DC (0.795). In contrast, MS and ALS have the worst recall results, with 0.697 and 0.429 in that order.

Table 7. NB Detailed Performance by Class.

TP Rate	Precision	Recall	F-Measure	MCC	PRC Area	Class
0.924	0.986	0.924	0.954	0.927	0.984	SLE
0.429	0.600	0.429	0.500	0.492	0.529	ALS
0.795	0.660	0.795	0.722	0.638	0.794	DC
0.697	0.852	0.697	0.767	0.731	0.752	MS
0.838	0.756	0.838	0.795	0.747	0.836	AIH
0.825	0.836	0.825	0.827	0.783	0.861	Weighted Average

Random Forest (RF). The Random Forest correctly classified 158 cases (79%) with a weighted average accuracy of 0.793 and an equivalent recall of 0.790. As shown in Table 8, metrics vary according to each class, emphasizing SLE and ALS. There is a significant difference between these two classes' F-score (SLE = 0.943 and ALS 0.308) and the MCC (SLE = 0.906 and ALS 0.285).

Table 8. RF Detailed Performance by Class.

TP Rate	Precision	Recall	F-Measure	MCC	PRC Area	Class
0.937	0.949	0.937	0.943	0.906	0.978	SLE
0.286	0.333	0.286	0.308	0.285	0.297	ALS
0.795	0.660	0.795	0.722	0.638	0.670	DC
0.545	0.750	0.545	0.632	0.582	0.722	MS
0.784	0.744	0.784	0.763	0.708	0.807	AIH
0.790	0.793	0.790	0.787	0.735	0.813	Weighted Average

Logistic Regression (LR). The Logistic Regression model had an accuracy of 84% (n = 168) with a weighted average precision of 0.852. According to Table 9, the general result of the classes of all metrics obtained values greater than 0.8.

The ALS class remains the one with the worst results among the others but showed an increase in recall (0.571), precision (0.667), F-score (0.615), MMC (0.604) and PRC Area (0.652) in comparison with the other classifiers.

Table 9. LR Detailed Performance by Class.

TP Rate	Precision	Recall	F-Measure	MCC	PRC Area	Class
0.924	1,000	0.924	0.961	0.938	0.984	SLE
0.571	0.667	0.571	0.615	0.604	0.652	ALS
0.864	0.704	0.864	0.776	0.710	0.836	DC
0.758	0.862	0.758	0.806	0.773	0.752	MS
0.757	0.737	0.757	0.747	0.688	0.831	AIH
0.840	0.852	0.840	0.843	0.803	0.873	Weighted Average

Multilayer Perceptron (MLP). The neural network correctly classified 156 instances (78%) with an average precision between classes equivalent to 0.778. In addition, an average F-score of 0.777 and an average PRC Area of 0.805 was obtained.

Furthermore, as with all other classifiers, SLE was the most accurately predicted class (0.944), while ALS was the lowest (0.333). The other quality metrics' specific values for each class are presented in Table 10.

Table 10. MLP Detailed Performance by Class.

TP Rate	Precision	Recall	F-Measure	MCC	PRC Area	Class
0.924	0.948	0.924	0.936	0.895	0.980	SLE
0.143	0.333	0.143	0.200	0.200	0.220	ALS
0.727	0.615	0.727	0.667	0.566	0.675	DC
0.636	0.677	0.636	0.656	0.591	0.660	MS
0.784	0.784	0.784	0.784	0.735	0.824	AIH
0.780	0.778	0.780	0.777	0.719	0.805	Weighted Average

4.4 Comparison between Models

Training Set. Table 11 compares the algorithms used based on the weighted average of the quality and accuracy metrics achieved in each model. In this scenario, the KNN-7 was chosen as the model with the best performance for this classifier.

Visualizing that the results obtained between the classifiers are similar, not presenting significant divergences is possible. However, it is noted that Naive Bayes is the only model that gives results above 0.8 for all metrics, except for MCC, in addition to having the highest recall (0.817). On the other hand, the KNN-7 presents the worst results in general among the classifiers.

From the accuracy, sensitivity, and F-score perspective, NB, LR, and SVM performed best in that order. As for the MCC, no model achieved a value equal to or greater than 0.8. However, all except the SVM denoted a PRC Area greater than 0.8.

Table 11. Comparison between Training Models by Detailed Average Performance.

TP Rate	Precision	Recall	F-Measure	MCC	PRC Area	Classifier
0.774	0.778	0.774	0.765	0.705	0.838	KNN-7
0.783	0.790	0.783	0.783	0.727	0.728	SVM
0.809	0.817	0.817	0.809	0.761	0.878	NB
0.781	0.783	0.781	0.780	0.722	0.826	RF
0.796	0.800	0.796	0.797	0.744	0.860	LR
0.780	0.778	0.780	0.777	0.719	0.805	MLP

Test Set. Table 12 compares each of the algorithms used, with the choice of KNN-7 as the best model for this classifier. The values present in each metric refer to the weighted average obtained in each model with the test set.

From this, it is observed that there are no such significant discrepancies between each classifier. Still, among the six metrics, Logistic Regression performed the best in all of them. In contrast, the KNN-7 remains the worst model, with lower results in 5 metrics, except for the PRC Area.

From the precision, recall, and F-score perspective, only NB and LR obtained results above 0.8. As for the PRC Area, all models are above 0.8, excluding the SVM.

Table 12. Comparison between Test Models by Detailed Average Performance.

TP Rate	Precision	Recall	F-Measure	MCC	PRC Area	Classifier
0.735	0.749	0.735	0.735	0.668	0.812	KNN-7
0.790	0.811	0.790	0.795	0.743	0.737	SVM
0.825	0.836	0.825	0.827	0.783	0.861	NB
0.790	0.793	0.790	0.787	0.735	0.813	RF
0.840	0.852	0.840	0.843	0.803	0.873	LR
0.780	0.778	0.780	0.777	0.719	0.805	MLP

Confusion Matrix. Table 13 depicts an interesting result plotted in the classes of Autoimmune Hepatitis and Crohn's Disease. The Logistic Regression model's confusion matrix during the training phase shows these two classes.

Table 13. LR Test Model Confusion Matrix.

	Predict Class	
Class	AIH	DC
AIH	28	9
DC	6	38

Best Classifiers. Based on the results obtained, verifying the performance of each model against the quality metrics and comparing them between classes and generalized weighted averages, it was possible to find the best models, as shown in Table 14.

Table 14. Best Models by Detailed Average Performance.

TP Rate	Precision	Recall	F-Measure	MCC	PRC Area	Classifier
0.840	0.852	0.840	0.843	0.803	0.873	LR
0.825	0.836	0.825	0.827	0.783	0.861	NB
0.790	0.811	0.790	0.795	0.743	0.737	SVM

5 Discussion

Regarding the applied models, based on the general average results between the classes for each classifier, it appears that the three best algorithms are Logistic Regression, Naive Bayes, and Support Vector Machine, as shown in metric LR presents the best results on all six measures; NB has the second-best result among the metrics, and SVM has the third-best result among the five metrics, with the worst result for the PRC Area.

However, even though LR and NB are models with excellent metrics, with NB being the best machine learning method among the others, it is necessary to look at the confusion matrix and the quality parameters for each class. During the training phase, both algorithms had difficulty predicting the ALS class during the training phase, with sensitivity below 50%. Furthermore, an individual in the ALS class was normally classified as in the MS class. Thus, of the total ALS (n = 43), 35% were classified as MS for LR and 42% for NB.

In contrast, for the SLE class, both NB and LR scored 299 cases (93%) during training and 73 cases (92.4%) in the test phase. The main reason for this difference, which causes many classification errors for ALS and hits for SLE, is the class imbalance. Looking at the overall sample set (n = 1000), SLE represents 40% of the total, while MS only 5%. Such imbalance is present as it resembles the real world, where ALS is a rare disease with low prevalence, and SLE is rare in South America but with a much higher prevalence. By having a larger number of cases with SLE in the dataset, the algorithms can understand the attributes of this class. In contrast, for ALS, which has few samples, it becomes more complex to train and identify the possible factors that determine this disease.

In addition to dataset imbalance, which makes the prediction process more difficult, the similarity of symptoms between diseases is another factor. All diseases share at least three symptoms among them, and the more they have in common, the more complex it is for the algorithm to classify based on these attributes alone.

DOI: 10.1007/978-3-031-36024-4_13

As seen in Table 13, it is noted that there is a tendency that when the classifier misses the DC class, it categorizes it as AIH and vice versa. These results denote a difficulty for the algorithm to distinguish these two classes depending on the scenario, in line with the literature, since AIH is a disease often associated concomitantly with other autoimmune diseases such as Crohn's Disease, Rheumatoid Arthritis, and thyroid diseases.

Finally, it is important to point out some caveats regarding this study. It is understood that a real database is a primordial point for the development of the work reliably and consistently with reality. Therefore, although the database has been created and supported by several references following logical criteria, it cannot fully portray the truth. Thus, it is recommended for future studies to search for a database of medical records of patients with rare diseases, such as the database from the Brazilian National Network of Rare Diseases (RARAS) [35], restricting it to autoimmune diseases and verifying which possible diagnoses are available.

Another relevant limiting factor concerns the sample set. As these are rare autoimmune diseases, this is a particular subset, making it difficult to find sufficient data for this type of analysis methodology. Furthermore, the lack of standardization among the data from medical records in such a small set is another point that hinders the use of certain attributes for the classification process. Even with these problems, there are articles described in the literature that address the standardization and structuring process for these types of data [36, 37].

Finally, this study uses only nine attributes (no class included), 3 of which are demographic data and 6 are symptoms. However, a patient's medical record has numerous attributes that can be considered during the analysis, not restricted to those of this study. Thus, it is interesting that these questions are raised in further studies, seeking the best possible methodology to work with these data types.

6 Conclusion

Given the results presented and the established validation metrics, it was possible to use machine learning models capable of predicting autoimmune disease, with Logistic Regression being the best general model and Naive Bayes the best machine learning classifier.

Both algorithms had satisfactory results regarding general quality parameters between classes. However, when considering the individual classes, none could accurately and reliably predict the ALS diagnosis. Furthermore, there is evidence of a slight but notable difficulty in distinguishing the DC and AIH classes. Therefore, the LR and NB models found can predict rare autoimmune diseases exclusively for this research. However, they are not maintained in real life due to the need for a feasible database, and they do not present any classification variant for diagnosing ALS.

Although a real dataset and medical records were not used, this work presents a promising approach to underpin the diagnosis of rare autoimmune diseases, supporting physicians in this challenging task. This methodology will be applied in future articles using real data from RARAS’s database and exploring different kinds of predictions, like the prognosis of patients with rare diseases, for example.

Acknowledgments

This study is supported by the National Council for Scientific and Technological Development (CNPq) process no. 44303/2019-7.

References

1. Gattorno, M., Martini, A.: Immunology and rheumatic diseases. In Textbook of Pediatric Rheumatology. 6th edn. W.B. Saunders, (2011).
2. Smith, D.A., et al.: Introduction to immunology and autoimmunity. Environmental Health Perspectives. 107, 661–665 (1999).
3. Bioemfoco: Doenças Autoimunes: Porque são chamadas assim e os avanços nas pesquisas, https://bioemfoco.com.br/noticia/doencas-autoimunes-avancos-pesquisas/, last accessed 2022/05/02.
4. Riedhammer, C., et al.: Antigen Presentation, Autoantigens, and Immune Regulation in Multiple Sclerosis and Other Autoimmune Diseases. Frontiers in Immunology. 6, (2015).
5. autoimmune association: Diagnosis Tips, https://autoimmune.org/resource-center/diagnosis-tips/, last accessed 2022/05/02.
6. Ministério da Saúde: Lúpus, https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/l/lupus, last accessed 2022/05/02.
7. Araújo, A.D., et al.: Expressões e sentidos do lúpus eritematoso sistêmico (LES). Estudos de Psicologia (Natal). 12, 119–127 (2007).
8. Borba, E.F., et al.: Consenso de lúpus eritematoso sistêmico. Revista Brasileira de Reumatologia. 48, 196–207 (2008).
9. Santos, S. de C.: Doença de Crohn : uma abordagem geral. Specialization dissertation, Universidade Federal do Paraná, Curitiba (2011).
10. Poli, D.D.: Impacto da raça e ancestralidade na apresentação e evolução da doença de Crohn no Brasil. Masters dissertation, Universidade de São Paulo, São Paulo (2007).
11. Stafford, I.S., et al.: A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. Digital Medicine. 3, (2020).
12. DeMarshall, C., et al.: Autoantibodies as diagnostic biomarkers for the detection and subtyping of multiple sclerosis. Journal of Neuroimmunology. 309, 51–57 (2017).
13. Saavedra, Y.B.: Análisis de Autoreactividad de Anticuerpos Leucémicos Soportado por Estrategias de Inteligencia Artificial.Doctoral dissertation, Universidad de Talca, Talca (2021).
14. Forbes, J.D., et al.: A comparative study of the gut microbiota in immune-mediated inflammatory diseases—does a common dysbiosis exist? Microbiome. 6, (2018).
15. Iwasawa, K., et al.: Dysbiosis of the salivary microbiota in pediatric-onset primary sclerosing cholangitis and its potential as a biomarker. Scientific Reports. 8, (2018).
16. Esclerose Múltipla (EM), https://www.einstein.br/doencas-sintomas/esclerose-multipla, last accessed 2022/05/04.
17. Chiò, A., et al.: Global Epidemiology of Amyotrophic Lateral Sclerosis: A Systematic Review of the Published Literature. Neuroepidemiology. 41, 118–130 (2013).
18. Correia, L. et al.: Hepatite autoimune: os critérios simplificados são menos sensíveis? GE Jornal Português de Gastrenterologia. 20, 145–152 (2013).
19. Pereira, A.B.C.N.G., et al.: Prevalence of multiple sclerosis in Brazil: A systematic review. Multiple Sclerosis and Related Disorders. 4, 572–579 (2015).
20. Pearce, F., et al.: Can prediction models in primary care enable earlier diagnosis of rare rheumatic diseases? Rheumatology. 57, 2065–2066 (2018).

21. Sociedade Brasileira de Reumatologia: Lúpus Eritematoso Sistêmico (LES), https://www.reumatologia.org.br/doencas-reumaticas/lupus-eritematoso-sistemico-les, last accessed 2022/11/10.
22. Souza, M.M. de., et al.: Perfil epidemiológico dos pacientes portadores de doença inflamatória intestinal do estado de Mato Grosso. Revista Brasileira de Coloproctologia. 28, 324–328 (2008).
23. Souza, M.H.L.P., et al.: Evolução da ocorrência (1980-1999) da doença de Crohn e da retocolite ulcerativa idiopática e análise das suas características clínicas em um hospital universitário do sudeste do Brasil. Arquivos de Gastroenterologia. 39, 98–105 (2002).
24. Tamega, A. de A., et al.: Grupos sanguíneos e lúpus eritematoso crônico discoide. Anais Brasileiros de Dermatologia. 84, 477–481 (2009).
25. Chen, J., et al.: Systematic review with meta-analysis: clinical manifestations and management of autoimmune hepatitis in the elderly. Alimentary Pharmacology & Therapeutics. 39, 117–124 (2013).
26. Schaefer, J., et al.: The use of machine learning in rare diseases: a scoping review. Orphanet Journal of Rare Diseases. 15, (2020).
27. Aslam, N., et al.: Multiple Sclerosis Diagnosis Using Machine Learning and Deep Learning: Challenges and Opportunities. Sensors. 22, 7856 (2022)
28. Myszczynska, M.A., et al.: Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. Nature Reviews Neurology. 16, 440–456 (2020)
29. Scikit Learn: Support Vector Machines, https://scikit-learn.org/stable/modules/svm.html, last accessed 2022/05/14.
30. IBM: K-Nearest Neighbors Algorithm, https://www.ibm.com/topics/knn, last accessed 2022/05/14.
31. Scikit Learn: Naive Bayes, https://scikit-learn.org/stable/modules/naive_bayes.html, last accessed 2022/06/05.
32. IBM, What is random forest, https://www.ibm.com/topics/random-forest, last accessed 2022/06/05
33. Scikit Learn: Linear Models, https://scikit-learn.org/stable/modules/linear_model.html, last accessed 2022/06/05.
34. Scikit Learn: Neural network models (supervised), https://scikit-learn.org/stable/modules/neural_networks_supervised.html, last accessed 2022/06/05.
35. Rede Nacional de Doenças Raras, https://raras.org.br/, last accessed 2022/11/10.
36. Alves, D., et al.: Mapping, infrastructure, and data analysis for the Brazilian network of rare diseases: protocol for the RARASnet observational cohort study. JMIR Res. Protoc. 10(1), e24826 (2021)
37. Yamada, D.B., et al.: National Network for Rare Diseases in Brazil: The Computational Infrastructure and Preliminary Results. In: Groen, D., de Mulatier, C. (eds) Computational Science – ICCS 2022. ICCS 2022. LNCS, vol 13352. Springer, Cham(2022).