MAY 23 2022

Neuro-symbolic interpretable AI for automatic COVID-19 patient-stratification based on standardised lung ultrasound data FREE

Leonardo Lucio Custode; Federico Mento; Sajjad Afrakhteh; ... et. al

Proc. Mtgs. Acoust 46, 020002 (2022)
https://doi.org/10.1121/2.0001600

Volume 46 http://acousticalsociety.org/

182nd Meeting of the Acoustical Society of America
Denver, Colorado
23-27 May 2022

Biomedical Acoustics: Paper 2pBAb6

Neuro-symbolic interpretable AI for automatic COVID-19 patient-stratification based on standardised lung ultrasound data

Leonardo Lucio Custode, Federico Mento and Sajjad Afrakhteh
Department of Information Engineering and Computer Science, University of Trento, Trento, TN, ITALY; leonardo.custode@unitn.it; federico.mento@unitn.it; sajjad.afrakhteh@unitn.it

Francesco Tursi
UOS Pneumologia di Codogno, ASST Lodi, Lodi, ITALY; francesco.tursi@asst-lodi.it

Andrea Smargiassi and Riccardo Inchingolo
Department of Medical and Surgical Sciences, Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, ITALY; andrea.smargiassi @policlinicogemelli.it; riccardo.inchingolo@policlinicogemelli.it

Tiziano Perrone
Department of Internal Medicine, IRCCS San Matteo, Pavia, ITALY; tiziano.perrone@gavazzeni.it

Giovanni Iacca and Libertario Demi
Department of Information Engineering and Computer Science, University of Trento, Trento, TN, ITALY; giovanni.iacca@unitn.it; libertario.demi@unitn.it

In the current pandemic, being able to efficiently stratify patients depending on their probability to develop a severe form of COVID-19 can improve the outcome of treatments and optimize the use of the available resources. To this end, recent studies proposed to use deep-networks to perform automatic stratification of COVID-19 patients based on lung ultrasound (LUS) data. In this work, we present a novel neuro-symbolic approach able to provide video-level predictions by aggregating results from frame-level analysis made by deep-networks. Specifically, a decision tree was trained, which provides direct access to the decision process and a high-level explainability. This approach was tested on 1808 LUS videos acquired from 100 patients diagnosed as COVID-19 positive by a RT-PCR swab test. Each video was scored by LUS experts according to a 4-level scoring system specifically developed for COVID-19. This information was utilised for both the training and testing of the algorithms. A five-folds cross-validation process was utilised to assess the performance of the presented approach and compare it with results achieved by deep-learning models alone. Results show that this novel approach achieves better performance (82% of mean prognostic agreement) than a threshold-based ensemble of deep-learning models (78% of mean prognostic agreement).

Published by the Acoustical Society of America

1. INTRODUCTION

The advent of COVID-19 has prompted researchers to look for a way to early detect and effectively monitor patients affected by SARS-Cov 2. In this context, the use of lung ultrasound (LUS) has been rapidly spreading. Due to its portability, cost-effectiveness, real-time imaging, and safety, LUS provided the possibility to be widely adopted to evaluate the state of lungs in patients affected by COVID-19.[1–9] LUS can also be used to monitor and triage symptomatic patients.[1] In particular, LUS is often exploited to detect COVID-19 associated interstitial pneumonia and follow its evolution.[2,10] To this end, different imaging protocols have been suggested together with semi-quantitative scoring systems.[11] In parallel, quantitative methods for assessing the lung parenchyma are being developed.[10,12–16] These methods, however, are not yet available for a wide deployment in emergency conditions, due to their current preliminary state. Thus, semi-quantitative scoring systems based on specific LUS imaging features (e.g., vertical and horizontal artifacts, or consolidations) have been widely used during the COVID-19 pandemic.[2] In this context, the use of artificial intelligence (AI) in semi-quantitative scoring systems can reduce the subjectivity in evaluation, as well as the evaluation time.[17–20]

In this study, we use data obtained following an imaging protocol based on 14 scan areas and a 4-level scoring system, which enables the classification of different lung states.[2] A recent study demonstrated the prognostic value of this methodology when assessing the cumulative score (i.e., the sum of scores in the above-mentioned 14 scan areas) at the exam level.[21] To evaluate our approach, we acquired 1808 LUS videos from 100 COVID-19 positive patients, which consist of 366,301 frames in total. These frames are then analyzed by to two DNNs[17] that were trained to do automatic scoring and segmentation of LUS frames based on the four-level scoring system.[2] In the next step, to switch from frame-level to video-level labeling, the output scores of these two DNNs are considered as features for automatic classification based on a Decision Tree (DT) induced by means of an evolutionary algorithm called grammatical evolution, following the approach similar to the one proposed in Ref. 22. Specifically, we compare the video-level scores given by our automatic approach with the scores given by expert clinicians. Indeed, to perform their evaluation, clinicians associate a score to each video rather than to each frame. We then assess the performance of our aggregation approach (both at video-level and exam-level) by comparing the results obtained by the proposed method with a previously reported empirical aggregation technique.[23] We make our data freely available for further developments[1].

This paper is organized as follows. We provide the dataset description and the details of the proposed approach in Section 2. The results of the proposed method and its comparison with previously reported results are presented in Section 3. Finally, our conclusions are reported in Section 4.

2. MATERIALS AND METHODS

In the following, firstly we introduce the summary of the datasets, and then we provide the details of our proposed approach.

[1]https://drive.google.com/drive/folders/1Or4dF2fAM23H5fd_yxtq1vyAS8b7pL0s

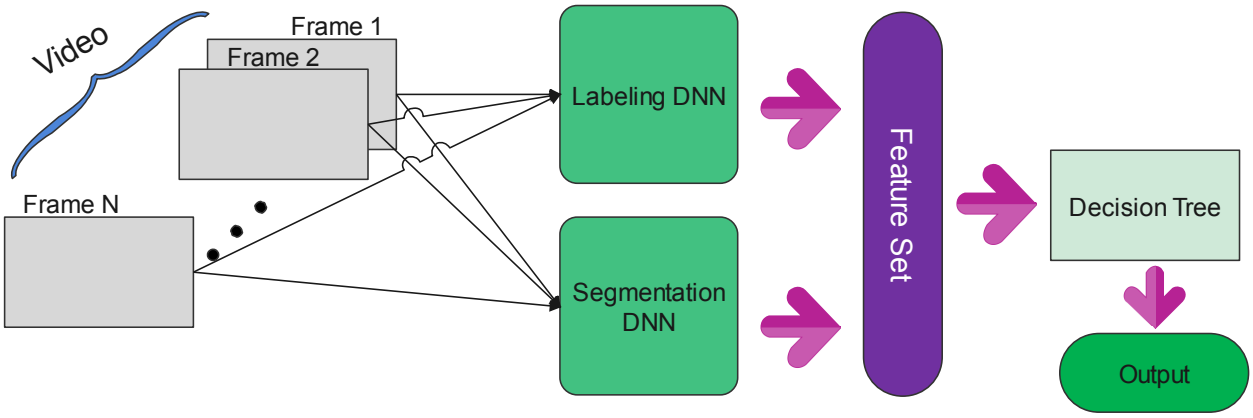

Figure 1: Block diagram of the prediction process.

A. DATASET

The dataset consists of 100 patients diagnosed as COVID-19 positive by a reverse transcription polymerase chain reaction (RT-PCR) swab test. From the 100 patients, 63 (35 male, 28 female; ages ranging from 26 to 92 years, and average age equal to 63.72 years) were examined within the Fondazione Policlinico San Matteo (Pavia, Italy), 19 (16 male, 3 female; ages ranging from 34 to 84 years, and average age equal to 63.95 years) within the Lodi General Hospital (Lodi, Italy), and 18 (8 male, 10 female; ages ranging from 23 to 95 years, and average age equal to 52.11 years) within the Fondazione Policlinico Universitario Agostino Gemelli (Rome, Italy). As a subgroup of patients was examined multiple times, on different dates (these patients needed to be monitored with LUS examinations on different dates, to track the development of the disease), a total of 133 LUS exams were performed (94 at Pavia, 20 at Lodi, and 19 at Rome). A total of 1808 LUS videos were thus acquired (1,290 at Pavia, 276 at Lodi, 242 at Rome), which consist of 366,301 frames (292,943 at Pavia, 44,288 at Lodi, 29,070 at Rome).

We divide the dataset into 5 different folds and allocate 4 folds to the training phase and 1 fold to the testing phase (to evaluate the generalizability of the proposed approach). The folds have been created in such a way that all the data related to a patient (even if there is more than one exam) lies in a single fold, to avoid "leakage" or biases in the folds.

B. STRUCTURE OF THE PROPOSED APPROACH

Fig. 1 shows the structure of the method proposed in this paper. The process is as follows: at first, 1808 videos are applied as input to two DNNs, which are described in Ref. 17. The first DNN only performs frames labeling, while the second one performs frame segmentation. More specifically, the difference between the two networks consists in the fact that the labeling network returns a score for the whole frame, while the segmentation network returns a score for each pixel contained in the frame. In order to use the information coming from the segmentation network, we assign to the frame a score equal to the highest score among all of its pixels.

Examples of LUS patterns with the associated scores are shown in Fig. 2. This four-level scoring system was designed to grade the lung condition with a score ranging from 0 to 3.[2] Specifically, the scores are assigned by visually analyzing LUS images to detect specific LUS imaging patterns that are associated to each score. As an example, score 0 is associated with a continuous pleural

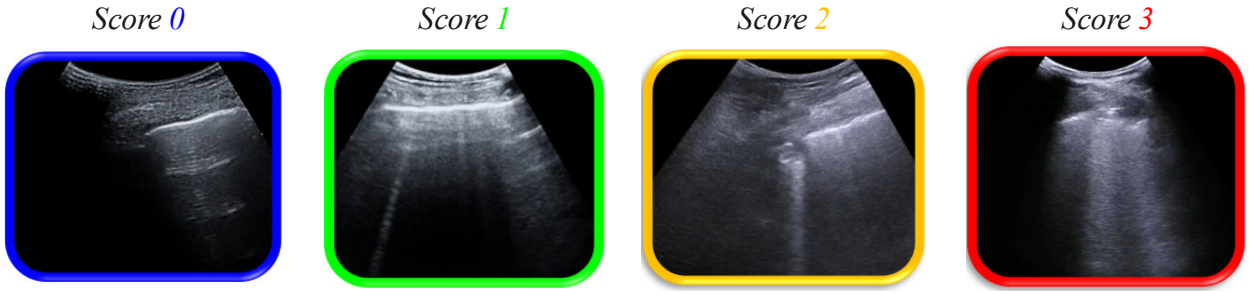

Figure 2: Some examples of frames with their corresponding scores.

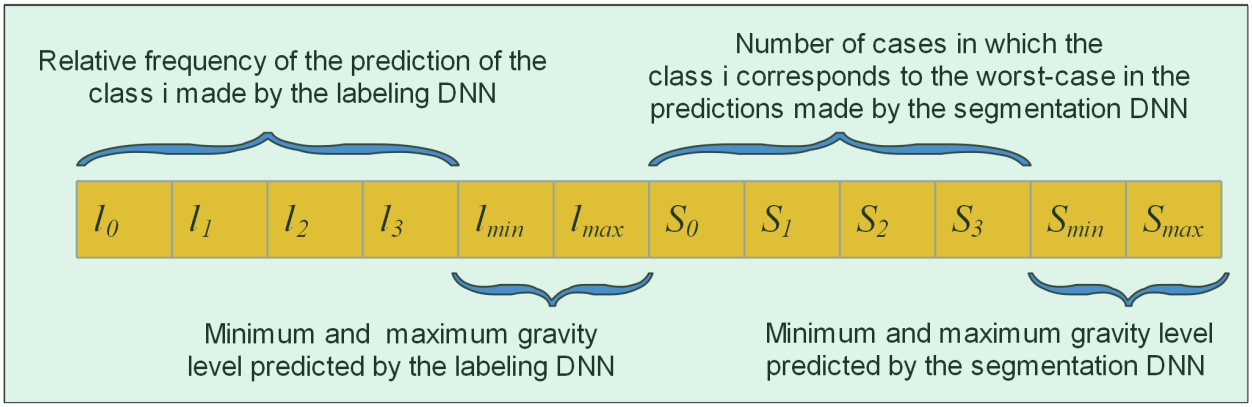

Figure 3: Final feature vector resulted from both DNNs in Fig. 1.

line and presence of horizontal artifacts, which are characteristic of a healthy lung surface.[2] In contrast, score 3 is associated with the presence of extended (>50% of the pleural line) vertical artifacts with or without large consolidations 2. Therefore, a higher score represents a worse state of the lung surface. The first DNN, taken from a spatial transformer, assigns one score for each frame. However, the second DNN, derived from U-Nets and DeepLab v3+, assigns multiple scores to each frame, depending on the segmented patters. For this network we consider the worst case (i.e., the highest score) found in each frame as the final score predicted. In the case that none of the relevant patterns occurs in a frame, we consider another Score (Score-1). In the feature extraction stage, to convert the features from frame-level to video-level, we aggregate the outputs of the two DNNs. For the labeling DNN, we simply use as features the relative frequency of the prediction of each class (i.e., the argmax of the output vector of each frame). For the segmentation DNN, instead, we aggregate the features by computing the relative frequency of the worst-case (i.e., maximum) predicted classes inside each frame. The resulting feature set is summarized in Fig. 3. Finally, in the classification phase, we feed this feature vector as input to a DT, for the final scoring of each LUS videos.

C. MULTI-OBJECTIVE GRAMMATICAL EVOLUTION OF DECISION TREES

We exploit the grammatical evolution (GE) algorithm[24] to evolve the structure and parameters of the aggregating DTs. GE is an evolutionary algorithm that allows the evolution of grammars, encoded in the Backus-Naur form. The grammar we used is shown in Table 1. By using this

grammar we can leverage GE to build decision trees that have good performance and that allow us to interpret the results returned by the neural networks.

Table 1: Grammar used to evolve the DTs. The symbol "\|" indicates the possibility to choose between different symbols.

Rule	Production
dt	$\langle if \rangle$
if	$if\ \langle condition \rangle\ then\ \langle output \rangle\ else\ \langle output \rangle$
condition	$\langle var \rangle \langle op \rangle \langle const \rangle \mid \langle var \rangle \langle op \rangle \langle var \rangle$
var	$\{input_i\};\ i \in [0, 12[$
op	$< \mid > \mid ==$
output	$0 \mid 1 \mid 2 \mid 3 \mid \langle if \rangle$
const	$[0, 1]$ with step 10^{-2}

Moreover, we set our DT optimization problem as a multi-objective one, by considering three objectives, namely the video-level, the exam-level, and the prognostic-level agreement, defined respectively as follows:

1. **Objective 1:** Video-level MSE (to be minimized):

$$\frac{1}{N_{videos}} \sum_{i=1}^{N_{videos}} (y_i^v - \hat{y}_i^v)^2; \tag{1}$$

2. **Objective 2:** Exam-level MSE (to be minimized):

$$\frac{1}{N_{exams}} \sum_{j=1}^{N_{exams}} (y_j^e - \hat{y}_j^e)^2; \tag{2}$$

3. **Objective 3:** Prognostic-level agreement (to be maximized):

$$\frac{1}{N_{exams}} \sum_{j=1}^{N_{exams}} \mathbb{I}(y_j^p = \hat{y}_j^p); \tag{3}$$

where y_i^v indicates the ground truth for video i; $y_j^e = \sum_{i=1}^{14} y_{j,i}^v$ is the ground truth for exam j, i.e., the sum of the scores of each video of that exam; $y_j^p = \mathbb{I}(y_j^e > 24)$ is the ground truth for the prognosis j, where $\mathbb{I}$ is the indicator function, that outputs 1 if the argument is true, otherwise 0. Note that the notation $\hat{y}_b^a$ refers to the approximation made by our system for the variable y_b^a. With these definitions, the multi-objective evaluation is implemented as shown in Alg. 1.

Algorithm 1: Multi-objective evaluation

Input: T: the DT to evaluate
Input: $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \mathbf{X}_4$: input features of each fold
Input: $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4$: video-level ground truth
Input: $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_4$: exam-level ground truth
Input: $\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_4$: prognostic-level ground truth
Result: f: a list of objective functions

$num\_folds \leftarrow 4$ `// Number of folds used for the training`
$e_v \leftarrow []$;
$e_e \leftarrow []$;
$a_p \leftarrow []$;
`// Iterate over the training folds`
for $i = 0; i < num\_folds; i++$ **do**
  `// For each fold, compute the metrics and concatenate`
  $e_v \leftarrow concatenate(e_v, [video\_mse(T, \mathbf{X_i}, \mathbf{v_i})])$ `// Eq. 1`
  $e_e \leftarrow concatenate(e_e, [exam\_mse(T, \mathbf{X_i}, \mathbf{e_i})])$ `// Eq. 2`
  $a_p \leftarrow concatenate(a_p, [prognostic\_agreement(T, \mathbf{X_i}, \mathbf{p_i})])$ `// Eq. 3`
end
`// Assign the worst-case to each metric`
$f \leftarrow []$;
$f[0] \leftarrow max(e_v)$;
$f[1] \leftarrow max(e_e)$;
$f[2] \leftarrow min(a_p)$;
return f;

3. RESULTS

We perform 10 independent runs of grammatical evolution with different random seeds. We consider for comparisons the best evolved DT (i.e., the DT that shows the maximum prognostic level across 10 different runs). In Table 2, we show, for each method with different thresholds (i.e., the ones used in Ref. 23) the descriptive statistics of the agreement of the method across the 5 folds (i.e., the four used for training, plus the one used for testing). The agreement for the three metrics is computed as follows:

1. Video-level agreement:

$$\frac{1}{N_{videos}} \sum_{i=1}^{N_{videos}} \mathbb{I}(|\, y_i^v - \hat{y}_i^v| \leq Th) \tag{4}$$

2. Exam-level agreement:

$$\frac{1}{N_{exams}} \sum_{i=1}^{N_{exams}} \mathbb{I}(|\, y_i^e - \hat{y}_i^e| \leq Th) \tag{5}$$

3. Prognostic-level agreement:

$$\frac{1}{N_{exams}} \sum_{i=1}^{N_{exams}} \mathbb{I}(y_i^p = \hat{y}_i^p) \tag{6}$$

where we use the same notation of Eq. 1-3 and Th indicates a given threshold.

As can be seen, the proposed method has been compared with the results presented in Ref. 23, which are are based on empiric combination of frame-level predictions made by DNNs. More specifically, in the "Labeling Only" method, only the labeling DNN have been used, while in "Labeling + Segmentation" method, the segmentation DNN has also been exploited in combination with the labeling DNN. The video-level agreement indicates the fraction of videos that are correctly classified by the algorithm, which is measured based on the degree of compliance between the experts' evaluation and the output of the algorithm. We denote the agreement characterized by an exact match between video-level scores as video-level agreement with threshold Th equal to 0, whereas we refer to the video-level agreement allowing a disagreement up to 1 point as video-level agreement with Th equal to 1. Therefore, the results of Table 2 for these two values of Th are presented separately. Similarly, in Table 3 and Table 4 we report the comparison between the results of our best evolved DT and the results reported in Ref. 23 for the case of exam-level and prognostic-level agreement respectively. The disagreement threshold Th is set to 2, 5 and 10 for exam-level case and 24 for the prognostic-level one. As the results of Tables 2-4 show, the proposed approach works better in most cases than the previous two methods presented in Ref. 23, considering all the three agreement levels, obtaining worst-case improvements of up to 5%. The best decision tree induced by our method is shown in Fig. 4.

Table 2: Descriptive statistics (in %) of video-level agreement on all the folds. "Th" refers to the threshold used as tolerance, whose values have been taken from Ref. 23.

Method	Th	Min	Mean	Std	Med	Max
Labeling Only[23]	0	44.70	**50.40**	4.28	50.68	**57.38**
	1	82.74	85.87	2.49	85.91	89.76
Labeling + Segmentation[23]	0	42.35	50.10	5.48	**51.43**	56.67
	1	82.74	85.87	2.43	85.36	89.29
Proposed approach	0	**45.88**	49.48	**2.24**	49.77	52.86
	1	**85.47**	**88.31**	**2.29**	**87.84**	**92.14**

4. CONCLUSIONS

Recently the evaluation of LUS images to assess the severity of COVID-19 patients has received significant attention. In addition, the use of artificial intelligence in combination with medical systems can provide greater flexibility and explainability. In this paper, we used two deep neural networks previously presented in the literature to extracted features (namely, prediction

Table 3: Descriptive statistics (in %) of exam-level agreement on all the folds. "Th" refers to the threshold used as tolerance, whose values have been taken from Ref. 23.

Method	**Th**	**Min**	**Mean**	**Std**	**Med**	**Max**
Labeling Only[23]	2	21.05	32.30	6.75	32.26	41.94
	5	45.00	56.95	**9.43**	57.90	70.97
	10	80.00	89.44	6.44	89.47	**100.00**
Labeling + Segmentation[23]	2	21.05	30.97	7.58	35.00	38.71
	5	**45.16**	59.44	13.23	52.63	**80.64**
	10	80.00	84.58	6.27	81.25	96.77
Proposed approach	2	21.05	**36.36**	**9.68**	**38.71**	**46.88**
	5	45.00	**63.98**	11.35	**62.50**	77.42
	10	**85.00**	**91.51**	**5.34**	**90.32**	**100.00**

Table 4: Descriptive statistics (in %) of prognostic-level agreement on all the folds. "Th" refers to the threshold used as tolerance, whose values have been taken from Ref. 23.

Method	**Th**	**Min**	**Mean**	**Std**	**Med**	**Max**
Labeling Only[23]	24	63.13	78.12	7.96	80.64	85.00
Labeling + Segmentation[23]	24	57.90	76.78	12.01	83.87	**90.00**
Proposed approach	24	**68.42**	**82.11**	**7.51**	**84.38**	**90.00**

scores) from LUS data, and then we aggregated those features by means of a decision tree (DT). To account for the multi-faceted nature of this form of aggregation, we optimized the DTs by means of multi-objective grammatical evolution, aimed at optimizing simultaneously frame-level, video-level and prognostic-level agreements. We then evaluated the results of the best evolved DT with respect to the three levels of agreement. Our results indicate that the performance of the proposed multi-objective approach is in most cases better than previously proposed methods based on empiric aggregation.

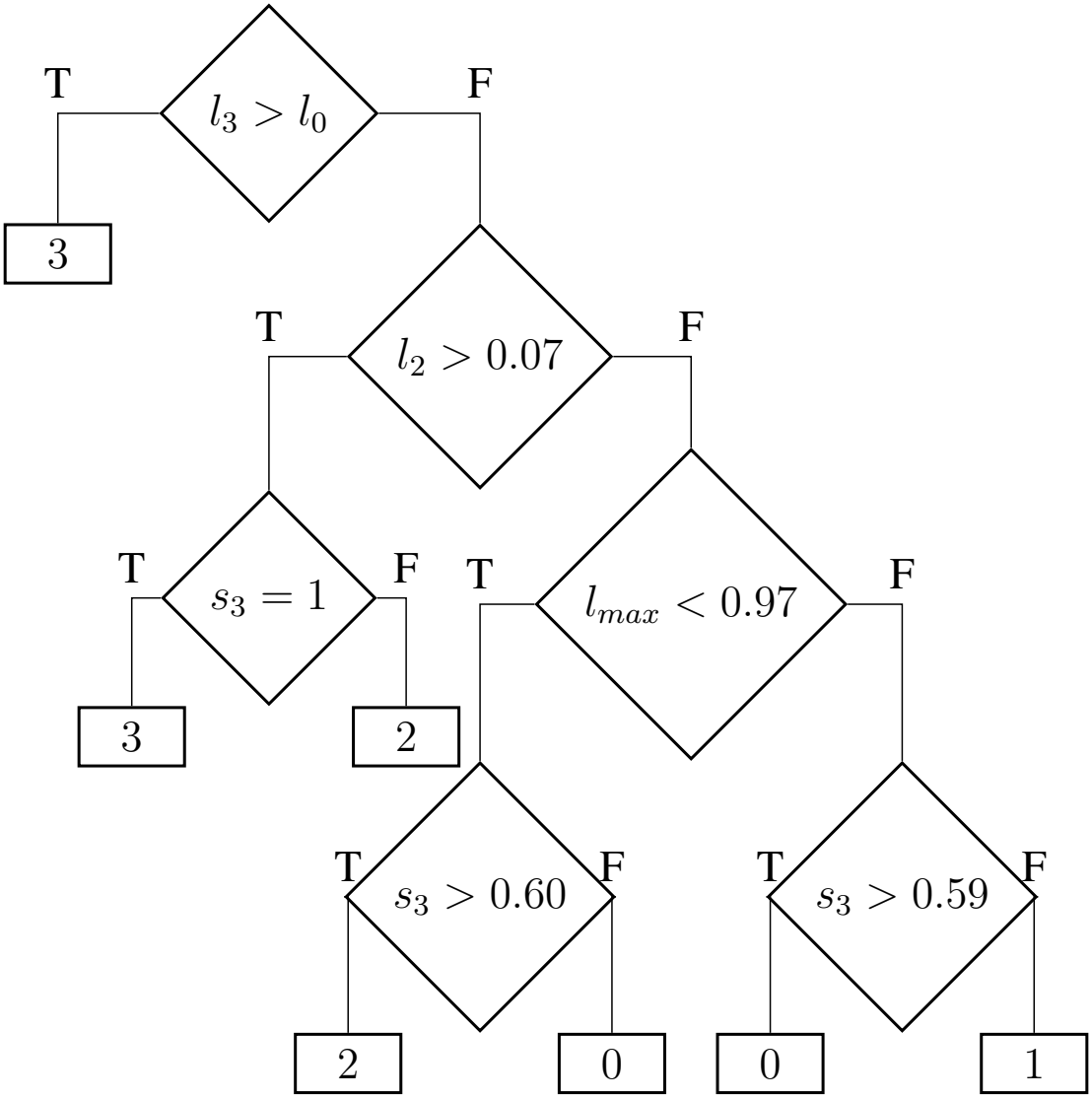

Figure 4: Best decision tree obtained with the proposed approach.

REFERENCES

[1] G. Soldati, A. Smargiassi, R. Inchingolo, D. Buonsenso, T. Perrone, D. F. Briganti, S. Perlini, E. Torri, A. Mariani, E. E. Mossolani, F. Tursi, F. Mento, and L. Demi, "Is There a Role for Lung Ultrasound During the COVID-19 Pandemic?," Journal of Ultrasound in Medicine **39**(7), 1459–1462 (2020).

[2] G. Soldati, A. Smargiassi, R. Inchingolo, D. Buonsenso, T. Perrone, D. F. Briganti, S. Perlini, E. Torri, A. Mariani, E. E. Mossolani, F. Tursi, F. Mento, and L. Demi, "Proposal for International Standardization of the Use of Lung Ultrasound for Patients With COVID-19," Journal of Ultrasound in Medicine **39**(7), 1413–1419 (2020).

[3] E. Poggiali, A. Dacrema, D. Bastoni, V. Tinelli, E. Demichele, P. Mateo Ramos, T. Marcianò, M. Silva, A. Vercelli, and A. Magnacavallo, "Can Lung US Help Critical Care Clinicians in the Early Diagnosis of Novel Coronavirus (COVID-19) Pneumonia?," Radiology **295**(3), E6–E6 (2020).

[4] P. Lomoro, F. Verde, F. Zerboni, I. Simonetti, C. Borghi, C. Fachinetti, A. Natalizi, and A. Martegani, "COVID-19 pneumonia manifestations at the admission on chest ultrasound, radiographs, and CT: single-center study and comprehensive radiologic literature review," European Journal of Radiology Open **7** (2020).

[5] A. Nouvenne, A. Ticinesi, A. Parise, B. Prati, M. Esposito, V. Cocchi, E. Crisafulli, A. Volpi, S. Rossi, E. G. Bignami, M. Baciarello, E. Brianti, M. Fabi, and T. Meschi, "Point-of-Care Chest Ultrasonography as a Diagnostic Resource for COVID-19 Outbreak in Nursing Homes," Journal of the American Medical Directors Association **21**(7), 919–923 (2020).

[6] K. Yasukawa and T. Minami, "Point-of-Care Lung Ultrasound Findings in Patients with COVID-19 Pneumonia," The American Journal of Tropical Medicine and Hygiene **102**(6), 1198–1202 (2020).

[7] C. Xing, Q. Li, H. Du, W. Kang, J. Lian, and L. Yuan, "Lung ultrasound findings in patients with COVID-19 pneumonia," Critical Care **24**(1), 174 (2020).

[8] Q.-Y. Peng, X.-T. Wang, L.-N. Zhang, and C. C. C. U. S. G. (CCUSG), "Findings of lung ultrasonography of novel corona virus pneumonia during the 2019–2020 epidemic," Intensive Care Medicine **46**(5), 849–850 (2020).

[9] G. Duclos, A. Lopez, M. Leone, and L. Zieleskiewicz, ""No dose" lung ultrasound correlation with "low dose" CT scan for early diagnosis of SARS-CoV-2 pneumonia," Intensive Care Medicine **46**(6), 1103–1104 (2020).

[10] L. Demi, "Lung ultrasound: The future ahead and the lessons learned from COVID-19," The Journal of the Acoustical Society of America **148**(4), 2146–2150 (2020).

[11] M. Allinovi, A. Parise, M. Giacalone, A. Amerio, M. Delsante, A. Odone, A. Franci, F. Gigliotti, S. Amadasi, D. Delmonte, N. Parri, and A. Mangia, "Lung Ultrasound May Support Diagnosis and Monitoring of COVID-19 Pneumonia," Ultrasound in Medicine and Biology **46**(11), 2908–2917 (2020).

[12] F. Mento, G. Soldati, R. Prediletto, M. Demi, and L. Demi, "Quantitative Lung Ultrasound Spectroscopy Applied to the Diagnosis of Pulmonary Fibrosis: The First Clinical Study," IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control **67**(11), 2265–2273 (2020).

[13] F. Mento and L. Demi, "On the Influence of Imaging Parameters on Lung Ultrasound B-line Artifacts, in vitro study," Journal of the Acoustical Society of America **148**(2), 975–983 (2020).

[14] G. Soldati and M. Demi, "The use of lung ultrasound images for the differential diagnosis of pulmonary and cardiac interstitial pathology," Journal of Ultrasound **20** (2017).

[15] K. Mohanty, J. Blackwell, T. Egan, and M. Muller, "Characterization of the Lung Parenchyma Using Ultrasound Multiple Scattering," Ultrasound in Medicine and Biology **43**(5), 993–1003 (2017).

[16] G. Zhang, J. Zhang, B. Wang, X. Zhu, Q. Wang, and S. Qiu, "Analysis of clinical characteristics and laboratory findings of 95 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a retrospective analysis," Respiratory research **21**(1), 74 (2020).

[17] S. Roy, W. Menapace, S. Oei, B. Luijten, E. Fini, C. Saltori, I. Huijben, N. Chennakeshava, F. Mento, A. Sentelli, E. Peschiera, R. Trevisan, G. Maschietto, E. Torri, R. Inchingolo, A. Smargiassi, G. Soldati, P. Rota, A. Passerini, R. J. G. V. Sloun, E. Ricci, and L. Demi, "Deep learning for classification and localization of COVID-19 markers in point-of-care lung ultrasound," IEEE Transactions on Medical Imaging **39**(8), 2676–2687 (2020).

[18] L. Carrer, E. Donini, D. Marinelli, M. Zanetti, F. Mento, E. Torri, A. Smargiassi, R. Inchingolo, G. Soldati, L. Demi, F. Bovolo, and L. Bruzzone, "Automatic Pleural Line Extraction and COVID-19 Scoring from Lung Ultrasound Data," IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control **67**(11), 2207–2217 (2020).

[19] W. Xue, C. Cao, J. Liu, Y. Duan, H. Cao, J. Wang, X. Tao, Z. Chen, M. Wu, J. Zhang, H. Sun, Y. Jin, X. Yang, R. Huang, F. Xiang, Y. Song, M. You, W. Zhang, L. Jiang, Z. Zhang, S. Kong, Y. Tian, L. Zhang, D. Ni, and M. Xie, "Modality alignment contrastive learning for severity assessment of COVID-19 from lung ultrasound and clinical information," Medical Image Analysis **69** (2021).

[20] O. Frank, N. Schipper, M. Vaturi, G. Soldati, A. Smargiassi, R. Inchingolo, E. Torri, T. Perrone, F. Mento, L. Demi, M. Galun, Y. C. Eldar, and S. Bagon, "Integrating Domain Knowledge into Deep Networks for Lung Ultrasound with Applications to COVID-19," IEEE Transactions on Medical Imaging 1 (2021).

[21] T. Perrone, G. Soldati, L. Padovini, A. Fiengo, G. Lettieri, U. Sabatini, G. Gori, F. Lepore, M. Garolfi, I. Palumbo, R. Inchingolo, A. Smargiassi, L. Demi, E. E. Mossolani, F. Tursi, C. Klersy, and A. Di Sabatino, "A New Lung Ultrasound Protocol Able to Predict Worsening in Patients Affected by Severe Acute Respiratory Syndrome Coronavirus 2 Pneumonia," Journal of Ultrasound in Medicine (2020).

[22] L. L. Custode and G. Iacca, "Evolutionary learning of interpretable decision trees" (2021).

[23] F. Mento, T. Perrone, A. Fiengo, A. Smargiassi, R. Inchingolo, G. Soldati, and L. Demi, "Deep learning applied to lung ultrasound videos for scoring COVID-19 patients: A multicenter study," The Journal of the Acoustical Society of America **149**(5), 3626–3634 (2021).

[24] C. Ryan, J. Collins, and M. O. Neill, "Grammatical evolution: Evolving programs for an arbitrary language," in *European Conference on Genetic Programming* (Springer, Berlin, Heidelberg, 1998), pp. 83–96.