Clinical Case Reports International

Short Communication
Published: 06 Dec, 2022

Neural-Symbolic System for Predicting COVID-19 Positivity

Fadja AN[1], Fraccaroli M[2] and Bizzarri A[2]*

[1]Department of Mathematics and Computer Science, University of Ferrara, Italy

[2]Department of Engineering, University of Ferrara, Italy

Abstract

Thanks to the huge amount of data collected by hospitals, it is now possible to exploit Machine Learning (ML) to build predictive models that can learn from data for identifying medical pathologies. The potential of Deep Learning (DL) and ML algorithms are well known but, in a field such as medicine, it is necessary to build interpretable and explainable systems instead of black-box systems as the de facto in DL. This work applies those techniques to both clinical data and Computed Tomography (CT) scans to predict COVID-19 positivity. To achieve an explainable model, we used both neural systems, for classifying and analyzing CT scans images, a symbolic model, Decision Tree, for analyzing clinical data concerning patients and a Neural-Symbolic architecture that integrates both systems using Hierarchical Probabilistic Logic Programming (HPLP). Experiments confirm that the proposed system provides a prediction accuracy of almost 90% and is able to provide explanation of the classifications.

Aim and Contributions

Recently, Artificial intelligence [1] and Machine Learning [2] have been widely used to assist and support physicians in managing the COVID-19 pandemic [3-7]. The pandemic locks almost all the whole world down and causes several deaths. The aim of this paper is to provide an explainable COVID-19 prediction system. Having an explanation as soon as a patient arrives at a hospital helps medical doctors to manage patients based on the cause of their infections. Our contribution is four-folds: We propose Random Forest (RF) [8,9] and Decison Tree (DT) [10] ML models, which predict whether a patient is COVID-19 positive from clinical data. Then, a CNN that classifies patients' lung, without lesions (class Normal) and with lesions (class Pneumonia), from CT scans. Finally, we propose a ML model that uses HPLP [11-15], an extension of Liftable probabilistic logic programming [16], to integrate the previous systems as shown in Figure 1, and predicts the COVID-19 positivity of patients arriving at the hospital. The integrated system is explainable i.e., it provides not only the prediction of COVID-19 positivity of patients but also a clear explanation of its decision. The explanation, which is rule-based, makes the system more interpretable and more trustful.

Experiments and Results

Three main experiments were performed: The first experiment predicts the probability of patients to be affected by COVID-19 from clinical data. The dataset was provided by Huazhong University of Science and Technology [17], Wuhan, China and consists of 1,521 patients of which 1,126 from Union Hospital (HUST-UH) and 395 from Liyuan Hospital (HUST-LH) (Figure 1). It includes 894 COVID-19 positive patients (COVID+) and 627 non-COVID-19 patients (COVID$^-$). All patients had 120 clinical attributes and 1,342 subjects had both CT and clinical data. Initially a RF is applied to extract the most relevant clinical attributes (also called features) that are important to identify patient's COVID-19 positivity.

This step is called feature importance extraction. The most relevant clinical attributes extracted are the following: Temperature, coefficient of variation of red blood cell volume distribution amplitude, standard deviation of red blood cell volume distribution amplitude, age, lymphocyte count, eosinophil count, eosinophil count, neutrophil count, hemoglobin and lymphocyte count percentage. After performing feature extraction, a new version of the dataset is created including only the clinical attributes deemed relevant by the RF. This dataset is used to train a DT and provided an accuracy similar to the RF's accuracy with the advantage that it is possible to extract

OPEN ACCESS

Correspondence:
Arnaud Nguembang Fadja, Department of Mathematics and Computer Science, University of Ferrara, Saragat 1, Ferrara, 44122, Italy,
E-mail: arnaud.nguembafadja@unife.it

Received Date: *14 Nov 2022*
Accepted Date: *30 Nov 2022*
Published Date: *06 Dec 2022*

Citation:
Fadja AN, Fraccaroli M, Bizzarri A. Neural-Symbolic System for Predicting COVID-19 Positivity. Clin Case Rep Int. 2022; 6: 1429.

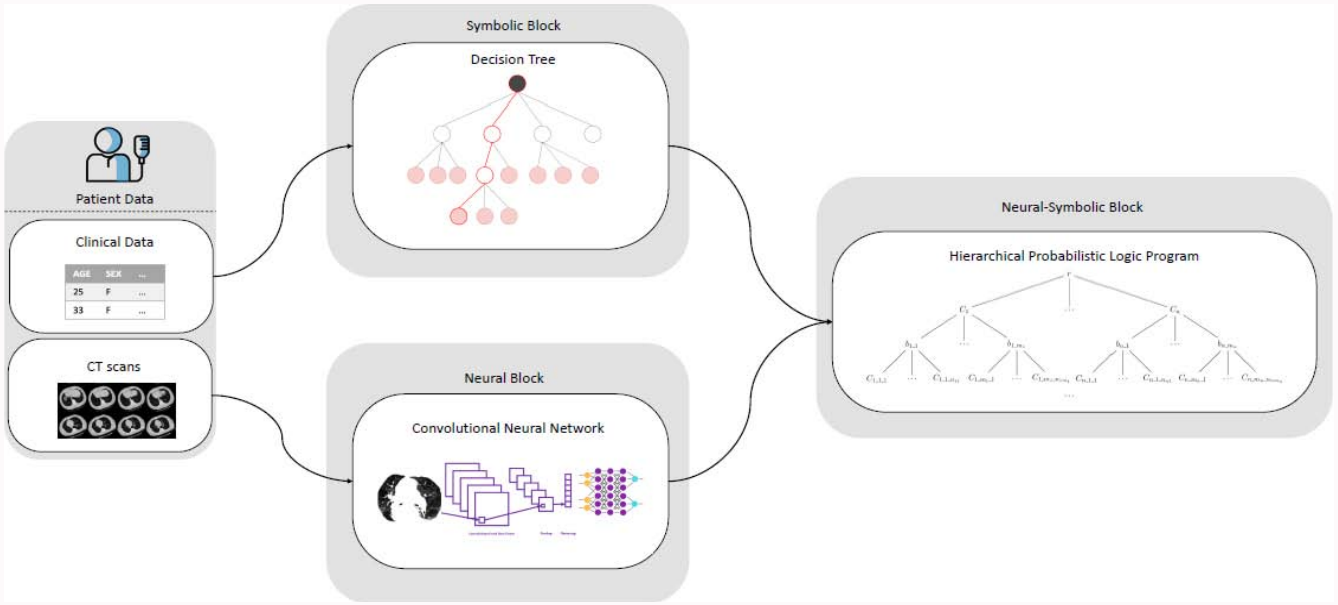

Figure 1: Neural-symbolic system.

Decision Paths (DPs). DPs are the conjunction of conditions starting from the DT root to the leaves. After training the DT, we obtained the following metrics on the set: An accuracy of 90.14%, an AUCROC of 0.9045 and an AUCPR of 0.9208.

The second experiment on CT scans uses 1006 patients with pneumonia and 336 patients with normal lungs. For each image, individual slices were extracted and preprocessed. The preprocessing phase consists of a segmentation phase to create a binary lung mask followed by the application of the mask to eliminate unnecessary parts of the images. Segmentation is done using the Hounsfield (HU) scale and, after applying the binary mask, the images were normalized. A total of 47,260 2D images were obtained for the training of a CNN. The images are grouped by the patient and divided into training (75%), validation (10%), and testing (15%). The test set includes 203 patients. The trained CNN is composed of the following parts: Four blocks composed of one convolutional layer with kernels of shape 3 × 3 and ReLU as activation function followed by a batch normalization layers with 64, 64, 128 and 256 neurons respectively.

These blocks are followed by a global average pooling layer, one fully connected layer with 512 neurons and one dropout layer. The output layer consists of 2 neurons associated with the two classes, normal and pneumonia lung. We achieved the followings results on the test set: An accuracy of 81.77%, an AUCROC of 0.823 and an AUCPR of 0.8709.

The last part of the experiment relies on HPLPs, a probabilistic logic program system [18], to integrate DT and CNN model. Given a target predicate to learn, patient COVID-19 positivity in our case, HPLP learns a program that consists of a set of logical rules annotated with probabilities. It learns from examples called interpretations which includes all clinical data concerning a patient. Two algorithms were experiments: HPLP_DEEP that applies gradient descent for learning and HPLP_EM that applies the expectation maximization algorithm. For more details on HPLPS see [11,14,15]. For learning we generated 203 interpretations, one for each patient in the test set. Each interpretation includes predicate on the decision path of the trained DT, the outputs of the DT ($COVID^+$ or $COVID^-$) and the CNN (normal/pneumonia lung). From the experiment, the following results are obtained: HPLP_DEEP achieved an AUCROC of 0.8188 and an AUCPR of 0.7210 while HPLP_EM achieved the better results with an AUCROC of 0.8956 and an AUCPR of 0.8144. Following an example of explainable COVID+ prediction.

References

1. Dick S. Artificial intelligence. HDSR. 2019.
2. Shalev-Shwartz S, Ben-David S. Understanding machine learning: From theory to algorithms. Cambridge University Press. 2014.
3. Cheng FY, Joshi H, Tandon P, Freeman R, Reich DL, Mazumdar M, et al. Using machine learning to predict ICU transfer in hospitalized COVID-19 patients. J Clin Med. 2020;9(6):1668.
4. Montomoli J, Romeo L, Moccia S, Bernardini M, Migliorelli L, Berardini D, et al. Machine learning using the extreme gradient boosting (XGBoost) algorithm predicts 5-day delta of sofa score at ICU admission in COVID-19 patients. JIM. 2021;1(2):110-6.
5. Roberts M, Driggs D, Thorpe M, Gilbey J, Yeung M, Ursprung S, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell. 2021;3(3):199-217.
6. Thomas MJ, Lal V, Baby AK, James A, Raj AK. Can technological advancements help to alleviate covid-19 pandemic? A review. J Biomed Inform. 2021;117:103787.
7. Fadja N, Fraccaroli M, Bizzarri A, Mazzuchelli G, Lamma E. Neural-symbolic ensemble learning for early-stage prediction of critical state of COVID-19 patients. Med Biol Eng Comput. 2022;60(12):3461-74.
8. Breiman L. Random forests. Machine learning. 2001;45(1):5-32.
9. Azar T, Elshazly HI, Hassanien AE, Elkorany AM. A random forest classifier for lymph diseases. Comput Methods Programs Biomed. 2014;113(2):465-73.
10. Quinlan JR. C4. 5: programs for machine learning. Elsevier. 2014.
11. Fadja N, Riguzzi F, Lamma E. Learning hierarchical probabilistic logic programs. Machine Learning. 2021;1-57.

12. Fadja N, Riguzzi F. Scalable probabilistic inductive logic programming for big data. 2020.

13. Fadja N, Riguzzi F, Lamma E. Learning the parameters of deep probabilistic logic programs. In PLP@ ILP. 2018:9-14.

14. Fadja N, Riguzzi F, Lamma E. Expectation maximization in deep probabilistic logic programming. In International Conference of the Italian Association for Artificial Intelligence. Springer. 2018:293-306.

15. Fadja N, Riguzzi F, Lamma E. Deep probabilistic logic programming. In PLP@ ILP. 2017:3-14.

16. Fadja N, Riguzzi F. Lifted discriminative learning of probabilistic logic programs. Machine Learning. 2019;108(7):1111-35.

17. Ning W, Lei S, Yang J, Cao Y, Jiang P, Yang Q, et al. Open resource of clinical data from patients with pneumonia for the prediction of COVID-19 outcomes *via* deep learning. Nat Biomed Eng. 2020;4(12):1197-207.

18. Fadja N, Riguzzi F. Probabilistic logic programming in action. In towards integrative machine learning and knowledge extraction. Springer. 2017:89-116.