Competition: COVID19 Detection in Blood Exams

Victor Henrique Alves Ribeiro
Artificial Intelligence Engineer
Hilab
Curitiba, Brazil
victor.ribeiro@hilab.com.br

Marcus Vinícius Mazega Figueredo
CEO
Hilab
Curitiba, Brazil
marcus@hilab.com.br

Abstract—The world currently suffers from the global COVID19 pandemic. Billions of people have been impacted, and millions of casualties have already occurred. Nevertheless, such numbers continue to increase as a new vaccine is still not made available. Therefore, it is of extreme importance to identify individuals who are or have been contaminated by the SARS-CoV-2 virus. Such an identification aids public health organizations and governments to plan actions to reduce the impacts of such a pandemic. In such a sense, Hilab is a remote laboratory company that performs dozens of types of blood exams, including serology tests for COVID19, where millions of exams have already been performed by the company in Brazil. To improve the detection of such a virus, machine learning methods can be used to aid laboratory experts into the decision making process. Therefore, this competition poses the difficult problem of building machine learning models with high confidence and accuracy for the detection of COVID19.

I. THE COMPETITION

Hilab is a remote laboratory company from Brazil which has thousands of blood scanner points distributed along the country, mostly on hospitals and pharmacies. The company has received great attention in the country due to its fast exam time, including for the detection of COVD19, in which the patients receive the exam's certificate in approximately 15 minutes. Once the blood sample is collected from the patient, it is digitalized and sent to a central laboratory in the city of Curitiba (Brazil), where expert biomedicians analyze the samples and indicate if the patient is infected.

Given the technology of the equipment, where blood samples are digitalized, and the high number of exams, enough data has become available to build strong machine learning models. Such models can aid the decision making process of the expert biomedicians, enabling a more accurate detection of blood infections. To improve the detection of COVID19 in individuals, Hilab proposes a competition on the difficult task of classifying different types of contamination using available processed data collected with the blood scanner and labeled by the expert biomedicians.

A. Dataset

Hilab will make available a single dataset, composed of thousands of labeled samples, which can be used by the competitors to train and validate their models. The dataset composes a multiclass classification problem, where several

Fig. 1. Hilab's blood sample scanner.

different classes must be correctly predicted given a multi-dimensional input. The dataset will be made available for the competitors in the following link: **available upon acceptance**.

B. Submission

The submission consists on a python package containing the source code for the preprocessing steps and the classification model. The whole solution can be composed of a single function, which receives the inputs (numpy 2D array) and returns the predicted outputs (numpy 1D array). Teams can be composed of a maximum of 5 individuals. Teams must submit their solutions, as well as their personal information, to **available upon acceptance**.

C. Evaluation Metrics

Solutions will be evaluated with a test dataset, which is not available for the competitors. The final score will be assessed using the Macro F_1 score, computed as follows:

$$Macro\ F_1 = \frac{1}{C} \sum_{c=1}^{C} \frac{2 \cdot TPR_c \cdot PPV_c}{TPR_c + PPV_c} \tag{1}$$

given

$$TPR_c = TP_c/(TP_c + FN_c) \qquad (2)$$
$$PPV_c = TP_c/(TP_c + FP_c) \qquad (3)$$

where TPR_c is the recall and PPV_c is the precision for class c.

D. Results

Final results with scores for the competitors will be made available during the competition's session at IJCNN, where the best solution (Highest F_1 score) will receive a prize of US$ 1,000.00 (value to be split among the team).