OPEN SOURCE IMAGING CONSORTIUM MEMBER CHALLENGE

Boehringer Ingelheim

Performance baselines with multi-modal data in predicting ILD diagnosis and mortality status

Philipp Schwarz, Plamen Danielov Petrov, Hao Li, Jan Seidel, Peide Li, Yi Liu, and Eric White
BOEHRINGER INGELHEIM PHARMA GMBH & CO. KG, INGELHEIM AM Rhein, Germany

OBJECTIVE

Self-supervised models have reported recent success in image recognition and in medical image classification. We studied the effectiveness of transferring pre-trained models with state-of-the-art architectures on anonymized HRCT scans and accompanying clinical data from the OSIC Data Repository to establish performance measure baselines for key tasks. Our experiments focus on the following problems: IPF identification, ILD diagnosis identification, and mortality status prediction.

METHODS

The baseline results are obtained via a two-step approach. Randomly selected slices from the HRCT scan data are transformed and loaded via a data pre-processing pipeline to a pre-trained backbone model which generates an intermediary embedding of the information in the scan. Next, the produced embedding is combined with key clinical characteristics and passed to a classification neural network head to compute final estimates. The classification head comprises three fully connected layers with appropriate regularization.

The transformation and loading pipeline include re-scaling and de-noising steps to ensure consistent scan orientation and patient lungs being in focus. The slices used for fine tuning the pre-trained network parameters were uniformly sampled from all available slices excluding the starting 25% and final 15%.

We selected 18 clinical variables with data availability in most patients and utilized in the classification model, which include age, sex, smoking history, exposure to risk factors, family history, AAB test information, and presence of symptoms.

The described framework was applied with calibrated adjustments to the three classification tasks. In the first task, conditions were encoded as IPF and non IPF. In the second task, conditions were encoded to one of five distinct classes — CHP, CTD-ILD, IPF, non-IPF/IIP, and others. In the third task, mortality status was investigated by a classification problem over three categories of cases - death within 1.5 years of observation, death after 1.5 years, but within 2.5 years of observation, and survival over 2.5 years of observation.

Performance on each task was evaluated via 5-fold cross-validation. At test time 20 scan slices were randomly selected for each patient and averaged estimates were derived for class predictions.

RESULTS

We establish baselines on all three tasks by measuring performance in terms of F1 score. We report benchmarks based on non-deep learning algorithms trained solely on clinical data and compare them to the results from transfer learning with a state-of-the-art backbone network based on HRCT scans, with or without the clinical information.

The choice of backbone network is key determinant of the performance of the proposed approach. We experimented with several backbone networks based on architectures from the ResNet and Efficient families and discovered the "efficientnet-b1" model to perform best on our set of problems. Therefore, we report only results utilizing that model. We also tested performance with backbone networks pre-trained with self-supervised contrastive learning strategy, which have reported recent successes on image classification tasks. We report outcomes from employing the BarlowTwins and SwAV models. All tested backbone networks were pretrained on the ImageNet dataset.

IPF IDENTIFICATION

Benchmark Linear Regression and Gradient Boosting models trained on clinical data achieved performance of above 0.88 on the identification of IPF. Models utilizing only scan images performed worse than the benchmark.

We enhance the baseline predictive power by including clinical data by including clinical data together with scan images and utilizing self-supervised contrastive learning strategy with BarlowTwins and SwAV models exceed F1 score of 0.91.

Benchmark: Clinical Data	
Method	*F1-Score*
Logistic Regression	0.885
Random Forest	0.855
GradientBoosting	0.883

Transfer Learning: Scan image classification	
Method	*F1-Score*
tf-efficientnet-b1-ns[1]	0.869
SwAV (Resnet-50)	0.873
BarlowTwins (Resnet-50)	0.877

Transfer Learning: Clinical Data + Scan image classification	
Method	*F1-Score*
tf-efficientnet-b1-ns	0.909
SwAV (Resnet-50)	0.913
BarlowTwins (Resnet-50)	0.913

1. tf-efficientnet-b1-ns: Weights of Tensorflow efficientnet-b1 noisy student

ILD DIAGNOSIS IDENTIFICATION

Benchmark Linear Regression model trained on clinical data achieved performance of above 0.84 on the identification of ILD diagnosis. Models utilizing only scan images performed worse than the benchmark.

We again improve upon predictive power by including clinical data together with scan images and utilizing self-supervised contrastive learning strategy with BarlowTwins and SwAV models, which result in an F1 score of 0.86. However, the performance gains in adopting this approach are smaller compared to the IPF identification task.

MORTALITY STATUS CLASSIFICATION

Benchmark: Clinical Data	
Method	*F1-Score*
Logistic Regression	0.840
Random Forest	0.809
GradientBoosting	0.827

Transfer Learning: Scan image classification	
Method	*F1-Score*
tf-efficientnet-b1-ns	0.755
SwAV (Resnet-50)	0.761
BarlowTwins (Resnet-50)	0.761

Transfer Learning: Clinical Data + Scan image classification	
Method	*F1-Score*
tf-efficientnet-b1-ns	0.855
SwAV (Resnet-50)	0.864
BarlowTwins (Resnet-50)	0.861

Benchmark Linear Regression model trained only on clinical data achieved performance of above 0.80 on the classification of disease progression status. Models utilizing only scan images performed worse than the benchmark and are not reported.

We again enhance the predictive power by including clinical data together with scan images and utilizing self-supervised contrastive learning strategy with BarlowTwins and SwAV models to exceed F1 score of 0.80.

In this task there are no performance gains in adopting a more complex approach.

Baseline: Clinical Data	
Method	*F1-Score*
Logistic Regression	0.804
Random Forest	0.741
GradientBoosting	0.792

Transfer Learning: Scan image classification	
Method	*F1-Score*
tf-efficientnet-b1-ns	—
SwAV (Resnet-50)	—
BarlowTwins (Resnet-50)	—

Transfer Learning: Clinical Data + Scan image classification	
Method	*F1-Score*
tf-efficientnet-b1-ns	0.800
SwAV (Resnet-50)	0.802
BarlowTwins (Resnet-50)	0.804

CONCLUSIONS

We use a two-step approach combining HRCT scan and clinical data based on transfer learning from a backbone network model with self-supervised pre-training on natural images to establish performance measure baselines for key tasks. We plan to use further self-supervised pre-training using not only ImageNet dataset but other unlabeled HRCT scans. Further, we will conduct more thorough hyperparameter searches. Future work will encompass the larger OSIC dataset to verify our initial results.