Improving clinical disease subtyping and future events prediction through a chest CT-based deep learning approach

Sumedha Singla[a)]
School of Computing and Information, University of Pittsburgh, Pittsburgh, PA 15213, USA

Mingming Gong
School of Mathematics and Statistics, The University of Melbourne, Parkville, VIC, Australia

Craig Riley
Chester County Hospital, University of Pennsylvania Health System, West Chester, PA, USA

Frank Sciurba
Department of Medicine, University of Pittsburgh Medical Center, Pittsburgh, PA 15213, USA

Kayhan Batmanghelich
Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15213, USA

(Received 2 August 2020; revised 30 October 2020; accepted for publication 9 December 2020; published 27 January 2021)

Purpose: To develop and evaluate a deep learning (DL) approach to extract rich information from high-resolution computed tomography (HRCT) of patients with chronic obstructive pulmonary disease (COPD).

Methods: We develop a DL-based model to learn a compact representation of a subject, which is predictive of COPD physiologic severity and other outcomes. Our DL model learned: (a) to extract informative regional image features from HRCT; (b) to adaptively weight these features and form an aggregate patient representation; and finally, (c) to predict several COPD outcomes. The adaptive weights correspond to the regional lung contribution to the disease. We evaluate the model on 10 300 participants from the COPDGene cohort.

Results: Our model was strongly predictive of spirometric obstruction ($r^2 =$ 0.67) and grouped 65.4% of subjects correctly and 89.1% within one stage of their GOLD severity stage. Our model achieved an accuracy of 41.7% and 52.8% in stratifying the population-based on centrilobular (5-grade) and paraseptal (3-grade) emphysema severity score, respectively. For predicting future exacerbation, combining subjects' representations from our model with their past exacerbation histories achieved an accuracy of 80.8% (area under the ROC curve of 0.73). For all-cause mortality, in Cox regression analysis, we outperformed the BODE index improving the concordance metric (ours: 0.61 vs BODE: 0.56).

Conclusions: Our model independently predicted spirometric obstruction, emphysema severity, exacerbation risk, and mortality from CT imaging alone. This method has potential applicability in both research and clinical practice. © *2020 The Authors. Medical Physics published by Wiley Periodicals LLC on behalf of American Association of Physicists in Medicine.* [https://doi.org/10.1002/mp.14673]

Abbreviations

DL	deep learning
HRCT	high-resolution computer tomography
CT	computer tomography
COPD	chronic obstructive pulmonary disease
r^2	the r-square coefficient of determination
GOLD	global initiative for obstructive lung disease
mMRC	the modified medical research council
ROC	receiver operating characteristic curve
BODE	the body mass index, obstruction, dyspnea and exercise capacity
HU	Hounsfield unit
LAA	low attenuation areas, lung voxels with Hounsfield Unit (HU) values <−950
SSDI	the social security death index \textbf{LFU:} longitudinal follow-up
CIP	Chest Imaging Platform
CNN	convolutional neural network
FEV1	forced expiratory volume in 1 s
FVC	forced vital capacity in 1 s
CLE	centrilobular emphysema
PH	proportional hazards
SD	standard deviation
ROS	random oversampling strategy
AUC-ROC	area under the receiver operating characteristic curve
AUC-PR	area under the precision recall curve
TP	true positive
FP	false positive
UMAP	Uniform Manifold Approximation and Projection
CI	confidence interval
LR	likelihood ratio
KM	Kaplan-Meier

1. INTRODUCTION

Chronic obstructive pulmonary disease (COPD) is characterized by persistent respiratory symptoms and irreversible airflow obstruction as measured by spirometry.[1] The measurement of spirometric obstruction, while traditionally used to define disease severity, is not sufficient to explain the many important dimensions required to fully characterize and manage COPD.[2] Airflow obstruction can be a result of varying combinations of emphysematous parenchymal destruction,[3] chronic airway remodeling,[4] and other poorly characterized imaging patterns, including fibrotic changes which are also common in smokers.[6] Hence, clinicians must adopt a comprehensive approach while assessing the patient with COPD, including identifying risk factors, standardized assessment of symptoms and comorbidities, estimating exacerbation risk,[7] and prognostication of survival. Other established tools for assessing symptoms are the modified Medical Research Council (mMRC) dyspnea scale and prognostication of survival using the body mass index, obstruction, dyspnea, and exercise capacity (BODE) index.[8,9] Although radiography has not been historically utilized in routine diagnosis or management of COPD, the increasing availability of computed tomography (CT) imaging from lung cancer screening programs provides a novel opportunity to leverage imaging data for improvement of patient care.

Much interest has been given to the use of CT imaging in subtyping COPD.[10] These efforts include assessment of specific features such as the percentage of low attenuation area (LAA),[11] blood vessel volume,[12] and airway counts.[13] Some of these methods rely on manual segmentation methods and are thus both labor-intensive and prone to operator error.[10,14–16] Recent promising work has incorporated texture-based feature extraction to identify COPD cases.[17–19] There are emerging works of using deep learning (DL) for COPD staging and subtyping.[20] However, most of the existing work concentrate on some aspect of COPD disease like only spirometry or only emphysema sub-typing. There is room for improvement to bring prediction of multiple patient-center outcomes to quantify COPD. Furthermore, much impact can be made by predicting patient's future exacerbation or survival, thus providing useful input to construct personalize treatment plans.

Our novel DL model followed a data-driven approach. It directly analyzed raw HRCT data and predicted clinical outcomes, without the need to manually segment or specify radiological features. Previous DL approach by Gonzalez et al.[20] processed slices (three orthogonal slices) of CT images and hence may not be able to characterize the volumetric impact of the disease. Our novel framework view each subject as a set of image patches from the lung region and thus analyzed the entire three-dimensional (3D) CT scan and required no image distortion due to resizing or cropping. Hence, our model can be trained on a modest GPU as we do not require a large memory to store a large field of view. Our model consists of three mutually dependent modules which regulate each other: (a) a *generative* network that extract local features from image patches and then reconstructs the image patch back from the latent features; (b) an *attention* mechanism that provides interpretability by adaptively weighting the patch-level features based on their contribution to overall prediction task; and (c) a *discriminative* network that aggregates the local features using adaptive weights, to form a patient representation and uses it to predict disease severity. In this work, we extended the prior model by Singla et al[21] to predict patient-relevant outcomes such as symptom scores, emphysema severity, and pattern, exacerbation risk, and mortality. When compared to the DL model by Gonzalez et al,[20] our method improved the prediction of important clinical variables, such as COPD disease severity and exacerbation risk. Furthermore, it can distinguish between centrilobular and paraseptal emphysema and can quantify the future risk of exacerbation based on the current CT image. The ability to estimate these clinically relevant features using only CT images has a potential application both to clinical care and research.

2. MATERIALS AND METHODS

2.A. Study cohort and imaging dataset

We evaluated our method on a dataset from the COPDGene study; an NIH funded multi-center clinical trial focused on the genetic epidemiology of COPD.[22] COPDGene includes 10 300 baseline participants, all of which were either current or former smokers. Each participant performed spirometry and had a high-resolution inspiratory and expiratory CT scan, using a standardized protocol.[22] The acquired CT scan images were assessed by trained experts to provide a visual quantification of the centrilobular and paraseptal emphysema severity. Survival information was collected using the social security death index (SSDI) search and the COPDGene longitudinal follow-up (LFU) program.

Our DL method is based on a convolutional neural network (CNN). A CNN is a type of artificial neural network used in image recognition that is specifically designed to process pixel data.[23] Convolutional neural network requires a fixed-size image as input. However, resizing the CT image alters the meaning of the density for each voxel. To avoid that, we represented each subject as a set of equally sized 3D patches. We extracted these patches from the parenchyma region in the chest. To achieve this, we first segmented the chest using chest imaging platform (CIP),[24] open-source software for quantitative CT imaging assessment. Next, we extracted 3D overlapping patches from parenchyma region of the chest. The number of patches depended on the volume of the lung that varied among individuals.

2.B. Deep learning architecture

The proposed model takes a set of volumetric image patches as input, that is, $\mathscr{X}_i = \{x_{ij}\}_{j=1}^{N_i}$, where N_i is the number of patches for patient i. The model learned to extract informative regional features from these patches x_{ij}, and then adaptively weight these features to form a fix-length representation for each patient. This patient representation is then used to predict disease severity (y_i). The general idea of our approach is shown in Fig. 1.

The first part of the model is: (a) a *generative* network that projects the raw image patch to a latent space and then reconstructs the image patch from the extracted latent features. The second part is (b) an *attention* network that learns a dynamic weight for each patch. The weight represents the relative importance of a given patch in making the patient-level predictions. The final part is (c) a *discriminative* network that aggregates the local information from patches in $\mathscr{X}_i$, based on their importance weights to create a patient-level representation and uses it to predict disease severity y_i. The model is trained end to end, by minimizing the below objective function:

$$\min_{\omega,\theta_e,\theta_d,\theta_a} \sum_i \mathscr{L}_d(y_i,\hat{y}_i(\mathscr{X}_i);\theta_e,\omega) + \lambda_1 \mathscr{L}_g\left(\mathscr{X}_i,\hat{\mathscr{X}}_i;\theta_e,\theta_d\right) + \lambda_2 \mathscr{R}(\mathscr{X}_i;\theta_e,\theta_a), \qquad (1)$$

where $\mathscr{L}_d(\cdot,\cdot)$ and $\mathscr{L}_g(\cdot,\cdot)$ are the discriminative and generative loss functions respectively and $\mathscr{R}(\cdot)$ is a regularization over the attention. The θ_e, θ_d, θ_a and ω are the parameters of each term. λ_1,λ_2 controls the balance between the terms. The sum is over number of subjects. Next, we discuss each term in more detail.

2.B.1. Generative network

The generative network is a convolutional autoencoder (CAE).[25] CAE consists of an encoder $\phi_e(\cdot)$, that extracts local image features from each patch $\left(i.e.,\phi_e(x_{ij};\theta_e)\in\mathbb{R}^d\right)$. These features are a summarization of the information in the raw image patch (or region) in a low dimensional "feature space." To regularize the feature extraction process, CAE have a decoder $\phi_d(\cdot)$. The decoder recovers the input patch back from the low-dimensional feature space as $\hat{x}_{ij} = \phi_d(\phi_e(x_{ij};\theta_e);\theta_d)$. In the absence of the decoder function, the feature extractor ϕ_e will be forced to retain only information that is sufficient for the underlying task of predicting y. As y is low dimensional as compared to d, ϕ_e will learn a highly redundant representation for each patch. To prevent this information loss, we regularize the auto-encoder using a distance loss defined as, $\mathscr{L}_g(\mathscr{X}_i,\hat{\mathscr{X}}_i;\theta_e,\theta_d) = \frac{1}{|\mathscr{X}_i|}\sum_{x_{ij}\in\mathscr{X}_i}\|x_{ij}-\hat{x}_{ij}\|_2$.

2.B.2. Attention network

The goal of the attention network is to learn a weight for each of the input image patches, such that the weight indicates the importance of a patch in predicting the overall disease severity of the lung. We used another neural network to learn these weights as $\boldsymbol{\alpha}_i = A(\phi_e(\mathscr{X}_i;\theta_e);\theta_a)$. We formulated

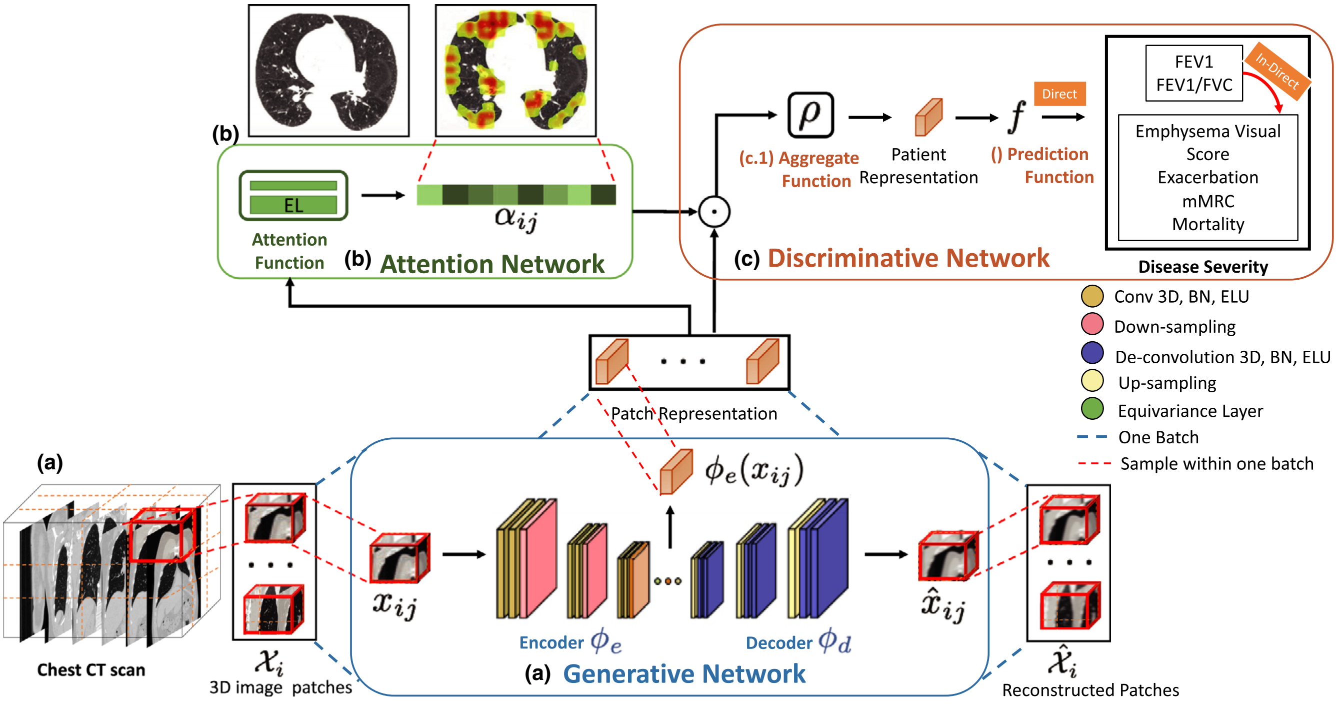

FIG. 1. The schematic of our model. (a) The input to our model is a three-dimensional (3D) computed tomography (CT) scan of the lung. The lung is divided into a set of equally sized, overlapping 3D image patches. (a) The **generative network** is a convolutional auto-encoder (CAE). The encoder function projects the raw image patch to a latent space and the decoder function reconstructs the image patch from the extracted latent features. (b) The **attention network** provides interpretability by weighting the patches based on their importance in predicting the disease severity. (c) The **discriminative network** (c.1) aggregates the local patch-level information information, based on their attention weights, to create a patient-level representation, and (c.2) uses it to predict disease severity. (b) An example of the weights learned by the adaptive weighting scheme overlaid on the input CT scan. Red color indicates higher relevance to the disease severity. In severe COPD cases, the red regions mostly focus on the bullae area, although not always. It also picks up normal regions because the absence of the normal tissue suggests more destruction by the disease and hence, more severe emphysema. Figure is best viewed in color.

the attention network $A(\cdot)$ as a feed-forward network, consisting of multiple equivariant layers (EL).[24] Assuming $\mathbf{H}_i \in \mathbb{R}^{N_i,d}$ where k^{th} row is $\phi(x_{ik};\theta_e) \in \mathbb{R}^d$, an equivariant layer is defined as

$$[\mathbf{H}_i]_k = \mathbf{W}\big([\mathbf{H}_i]_k - \max(\mathbf{H}_i, 1)\big) + \mathbf{b}, \tag{2}$$

where $[\mathbf{H}_i]_k$ denotes k^{th} row of $\mathbf{H}_i$ and $\max(\mathbf{H}_i, 1)$ is the max over rows. $\mathbf{W} \in \mathbb{R}^{L \times d}$, $\boldsymbol{b} \in \mathbb{R}^L$ are the parameters of the EL. Such formulation ensures that the weight of any patch depends not only on the corresponding patch feature but also on the features of all the other patches in a patient. Next, we pass the output of the EL layers to a softmax function, to obtain a distribution of weights over the patches. This ensures that the weights $(\boldsymbol{\alpha}_i)$ are non-negative numbers that sums to 1.

For better interpretability, the weight vector, $\boldsymbol{\alpha}_i$, should follow a sparse distribution. Increased sparsity pushes some weights terms, α_{ij}, to zero, and hence, it increases the interpretation by focusing on only the patches relevant for the prediction task. The best sparse constraint is to use the ℓ_0 norm over the weight vector, as it directly counts the number of non-zero elements. However, optimizing the ℓ_0 norm is problematic as it is not differentiable. ℓ_1 norm is usually used as a surrogate for ℓ_0 norm. In our formulation, the weights α_{ij}, have non-negative values that sum to 1 $(i.e., \|\boldsymbol{\alpha}_i\| = \sum_j \alpha_{ij} = 1)$. Hence, its derivative is zero, and using an ℓ_1 norm over the weight vector will not result in a sparse solution. To ensure high sparsity, we use a log-sum function as a regularizer. Minimizing $\sum_j \log \alpha_{ij}$ is equivalent of maximizing KL-divergence from the uniform distribution. The uniform distribution assigns the same weight to all the patches within one subject, that is, $\max_{\boldsymbol{\alpha}_i} \mathrm{KL}([\frac{1}{N_i}, \cdots, \frac{1}{N_i}], \boldsymbol{\alpha}_i) = \max_{\boldsymbol{\alpha}_i} \sum_j \frac{1}{N_i} \log \frac{1}{N_i} - \sum_j \frac{1}{N_i} \log \alpha_{ij} \equiv \min_{\alpha_i} \sum_j \log \alpha_{ij}$. We defined the regularization term as, $\mathscr{R}(\mathscr{X}_i;\theta_e,\theta_a) = \sum_{j=1}^{N_i} \log(\alpha_{ij} + \varepsilon)$ and add it to the loss function in Eq. (1).

2.B.3. Discriminative network

The discriminative network predicts the disease severity using the patient representation as

$$\hat{y}_i(\mathscr{X}_i) = f(\rho(\phi_e(\mathscr{X}_i, \theta_e)), \omega). \tag{3}$$

The discriminative network takes the patch-level features $\big(i.e., \phi_e(x_{ij};\theta_e)\big)$ extracted by the encoder as input. It transforms the patch-level features using composition of two functions: (a) The aggregate function $\rho(\cdot)$. It is a permutation invariant function that aggregates the patch-level features to form a fixed length patient representation. (b) A prediction function $f(\cdot;\omega)$, parameterized by ω. It takes the patient representation extracted by $\rho(\cdot)$ as input, and estimates the disease severity. Finally, $\mathscr{L}_d(y_i, \hat{y}_i(\mathscr{X}_i);\theta_e, \omega)$ is a regression or classification loss function between predicted and true value.

Conceptually, the aggregate function makes the prediction of disease severity less sensitive to the precise location within an image. It does so by aggregating the information from the local patches. One possible formulation of aggregate function is maximum function defined as, $\rho(\cdot) = \max(\phi_e(x_{i1}), \cdots, \phi_e(x_{iN_i}))$. The maximum function chooses only the highest value in each of the latent dimensions. $\max(\cdot)$ is not sensitive to values of the arguments that are less than the maximum value, that is, max({10,1,1,1}) = max({10,9,9,9}). Hence, its gradient is independent of most of the changes in the latent space. Another choice for aggregate function is an average function, defined as $\rho(\cdot) = \frac{1}{N_i}\sum_{j=1}^{N_i} \phi_e(x_{ij})$. It considers all the feature values and hence, spread out the volume of the latent space evenly. The average function assumes an equal contribution of all the local patches towards final disease severity. However, COPD disease is often attributed to the diffuse air-sacks obstruction spread unevenly throughout the lung. To incorporate the disease's diffused effect, we adaptively weight the patch-level features to create the patient representation as, $\rho(\cdot) = \sum_{j=1}^{N_i} \alpha_{ij} \phi_e(x_{ij})$. An attention network, described in Section 1, learns the weights (α_{ij}). This formulation also helps in interpretability, as the final weights highlight the regions based on their contribution to the prediction, as shown in Fig. 1(b).

The architecture of the encoder function consists of stacked convolutional layers which down-sampled the patches while doubling the number of channels. The decoder function consists of transposed convolutional layer (or deconvolutional layer) which upsample the features while cutting the number of channels to half. Each convolutional layer employs batch normalization for regularization, followed by an exponential linear unit (ELU)[26] for nonlinearity. The attention network has two equivalence layers with sigmoid activation function, followed by a softmax layer. The model is trained using an Adam optimizer[27] with hyperparameters $\beta_1 = 0$ and $\beta_2 = 0.999$ and a fixed learning rate of 0.001. The dimension of the feature vector is 128. The trade-off hyperparameters are $\lambda_1 = 10$ and $\lambda_2 = 1$. The experiments are performed on two NVIDIA p100 GPUs, each with 16GB GPU memory. The source code is available at https://github.com/batmanlab/Subject2Vec. The detailed architecture can be found in the Supplementary Material.

In our analysis, we used full-inspiration CT images, which were resampled to isotropic 1 mm^3. We worked on the fixed range of intensity values between −1024 and 240 HU, as suggested by Bhalla et al.[5] The number of patches in a subject (N_i) may vary between subjects. A large patch size or a high overlap between the patches increases the N_i for a subject. All the patches of a subject must be processed in the same batch, as they are required to learn the patient representation, which is then used to predict the disease severity. The available memory in a GPU memory restricts the maximum number of patches that can be processed in a single batch. We experimented with different values and finally used a patch size of 32 × 32 × 32 with a 40% overlap and an upper limit of 1000 patches per batch in our experiments. The average *Ni* for this setting is 700 patches per subject. We consider one subject per batch as shown in Fig. 1.

2.C. Setup for experiments

We presented an analysis of the performance of our model for predicting patient-centered outcomes related to COPD. We trained two versions; (a) **Direct**: the model was trained to predict the forced expiratory volume in 1 s (FEV1) and the FEV1/forced vital capacity (FVC) ratio, along with a clinical outcome of interest to represent disease severity. We separately trained one such model for each of the target outcomes. (b) **Indirect**: the model was trained only once to predict FEV1 and FEV1/FVC as disease severity. The patient representations from such model were then used in a separate regression analysis to predict other clinical outcomes of interest. The idea is to learn generalized patient representations by training the model for one clinical variable (spirometry) and testing on another clinical output (emphysema score) which the models have not seen previously. If two clinical variables are correlated, we should be able to capture much variance. Of course, training directly for the clinical variable, as in direct version, will achieve better results. For all results, we reported average test performance in five-fold cross-validation. We compared the performance of our method against

1. Baseline: The low attenuation area (LAA) features. **LAA-950** is defined as the total percentage of both lungs with attenuation values less than −950 Hounsfield units on inspiratory images. LAA-950 signifies radiographic emphysema.[11]
2. The **nonparametric** method proposed by Schabdac et al.[19] In this method, handcrafted image features were extracted for each patient, and nonparametric density estimation was performed to assign a characteristic vector to each patient.
3. The classical **k-means** algorithm applied to image features extracted from local lung regions.[19] A similar approach was suggested by Ash et al.[28]
4. The previous state-of-the-art method based on **CNN** also, applied to the COPDGene.[20]

For the first three methods, we reproduced the methods and reported results based on our experiments. For the last method, we reuse the numbers reported by the authors. The COPD outcomes used in our experiments are summarized in Table I.

2.C.1. Spirometry measures

As part of the pulmonary function test, following spirometry values were evaluated for all the participants in COPDGene: the forced expiratory volume in 1 s (FEV1) and the FEV1/forced vital capacity (FVC) ratio. All spirometric values were expressed as percentage of predicted values. Participants were classified as obstructed or non-obstructed under the 2019 Global Initiative for Chronic Obstructive Lung Disease (GOLD) guidelines using a fixed FEV1/FVC ratio of 0.7.[1] We defined the disease severity as the GOLD stages of 0 (nonobstructed) through 4 (very severely obstructed). Following the GOLD guidelines, in our experiments, we first train the model to predicted FEV1 and FEV1/FVC ratio, and then use these values to diagnose and stage COPD.

2.C.2. Visual emphysema score

In the COPDGene cohort, radiographic centrilobular (CLE) and paraseptal emphysema were scored on inspiratory scans by a trained research analysts using the Fleischner Society classification system. Detailed methods for emphysema visual quantification are provided by Lynch et al.[14] They grade the severity of CLE parenchymal emphysema on a scale of zero to five using labels: none, trace, mild, moderate, confluent, and advanced destructive emphysema. While

TABLE I. Summarization of the clinical outcomes considered in the experiments and their numerical type and values.

Clinical outcomes	Type	Values	Description
Spirometry measures			
FEV1	Continuous		Percentage predicted forced expiratory volume in 1 s
FEV1/FVC	Continuous		FEV1 ratio with forced vital capacity (FVC)
COPD	Binary	0 or 1	True if FEV1/FVC > 0.7
GOLD stages	Categorical	0–4	The GOLD stages of 0 (non-obstructed) through 4 (severely obstructed).
Visual emphysema score			
Centrilobular emphysema (CLE)	Categorical	0–5	CLE parenchymal emphysema severity score using values, none (0) to advanced destructive emphysema (6).
Paraseptal emphysema	Categorical	0–2	Specified using three labels: none, mild, and substantial.
Acute exacerbation			
Historic exacerbation	Binary	0 or 1	True if patient have experienced exacerbation in the last 1 yr.
Future exacerbation	Binary	0 or 1	True if patient reported experiencing an exacerbation by the 5th yr followup.
Others			
mMRC dyspnea scale	Categorical	0–4	The modified Medical Research Council (mMRC) dyspnea scale
Mortality	Binary	0 or 1	Vital status

paraseptal emphysema was scored using three labels: none, mild, and substantial.

2.C.3. Acute exacerbations

In the COPDGene study, the exacerbations of COPD were self-reported and were quantified by the subject recall on questionnaires. A participant recorded a positive experience of an acute exacerbation if, in the last year, they had experienced at least one episode of increased dyspnea, cough, or sputum production, resulting in admission to the hospital or changing of their treatment plan. Approximately 20% of the subjects reported experiencing at least one exacerbation before enrolling in the study. We used the HRCT acquired at the baseline visit to predict both historical and future exacerbations. The future exacerbation prediction used exacerbations reported by the longitudinal follow-up participants at the subsequent 5-yr follow-up visit.

2.C.4. mMRC dyspnea scale

Subjects completed the mMRC dyspnea scale during their baseline visit. The scale ranges from 0 (dyspnea only with strenuous exertion) to 4 (dyspnea with activities of daily living) and is used to guide therapeutic strategies in patients with COPD.[29,30]

2.C.5. Mortality

We used the vital status and censoring time information provided in the mortality dataset to perform survival analysis. In the COPDGene cohort, the mean time between phase 1 data and the censoring time is approximately five years. Nearly 13% of subjects were reported deceased either in the SSDI search or in the COPDGene LFU. We used Cox proportional hazards (PH) model[31] to predict survival utilizing the probability of death predicted by patient representation against age, gender, smoking status and center of enrollment as fixed covariates. Next, we used Kaplan–Meier plots stratified by quantile of predicted probabilities of death to visualize the results. The Kaplan-Meier plot shows the probability of survival plotted against time. We tested the PH assumption by performing a correlation between each of the covariates and their corresponding set of scaled Schoenfeld residuals with time.[32] A nonsignificant p-value for this test supported the PH assumption. In another test, we checked the global statistical significance of the Cox model. The test validated the null hypothesis that the variables have no association with survival. If the test failed to reject the null hypothesis, this would suggest that removing the variables from the model will not substantially harm the fit of that model. This global test is performed using three alternative tests: the likelihood-ratio test, the Wald test, and the score log-rank statistic. The survival analysis was performed using the lifelines library in Python[33] and the survival package in R.[34] We also compared the performance of our survival model against the univariate Cox regression model using intensity features (LAA-950) and the BODE index. The multidimensional BODE index has been shown to predict survival in cohort studies of COPD.[9] For the Cox PH model, we reported the results in terms of concordance, which is like the AUC-ROC statistic in binary classification.

2.D. Statistical analysis

Data for continuous variables are presented as mean ± standard deviation (SD). The percentage predicted FEV1 and FEV1/FVC ratio were evaluated as continuous variables. Exacerbation risk was defined as a binary variable, where subjects with one or more respiratory exacerbations were considered positive. Survival over 5-yr period was also encoded as a binary variable. The multicategory emphysema visual score and mMRC symptom score were defined as categorical variables.

We performed regression analysis for continuous variables and reported the performance in terms of the r-square coefficient of determination (r^2). We used a logistic regression model for binary variables. To make binary regression robust to class imbalances, we performed a random oversampling strategy (ROS). In ROS, we increased the number of instances in the minority class by randomly replicating them. Thus, ROS prevents the decision function from favoring the majority class. For binary classification, we reported area under the receiver operating characteristic curve (AUC-ROC), the area under the precision-recall curve (AUC-PR), and the recall. PR-AUC and recall statistics provide a better view of the classifier's performance in identifying subjects belonging to a minority class.[35] To calibrate the confidence of the classifier, we used the Hosmer–Lemeshow calibration test.[36] We subgrouped the subjects into ten risk-groups based on their predicted probabilities. We visualized the results in a calibration plot with predicted risk plotted against the observed risk for each subgroup.

We used multiclass ordinal regression for the categorical variable. The ordinal categories captured the level of disease progression (from mild to severe) in the subjects. Following the ordinal classification approach provided in Ref. [37] we trained our model by transforming the k-class ordinal regression problem to $k-1$ binary classification problems. We reported the classification accuracy and the percentage of the times the predicted class laid within one class of true value (*one-off*). To test whether the predicted classification probabilities are correctly calibrated, we use the Hosmer-Lemeshow[36] calibration test.

3. RESULTS

3.A. Spirometry measures

Our model attained an r^2 of 0.67 ± 0.03 for the FEV1 and 0.74 ± 0.01 for the FEV1/FVC ratio, which is significantly better than previously reported approaches (see Table II,

TABLE II. Results for predicting spirometry measurements and using them to diagnose and stage COPD.

			COPD diagnosis[b]			GOLD[c]	
Method	FEV1 R-square	FEV1/FVC R-square	AUC ROC[d]	AUC PR[e]	Recall	% accuracy	% accuracy *one-off*
Ours (direct)[a]	**0.67 ± 0.03**	**0.74 ± 0.01**	0.82	**0.72**	**0.80**	**65.44**	**89.14**
CNN[20a]	0.53	–	**0.86**	–	–	51.10	74.90
Non-parametric[19a]	0.58 ± 0.03	0.70 ± 0.02	0.79	0.70	**0.80**	58.85	84.15
K-Means[a]	0.56 ± 0.01	0.68 ± 0.02	0.77	0.68	**0.81**	57.27	82.28
LAA-950[a]	0.45 ± 0.02	0.60 ± 0.01	0.75	0.64	0.70	55.75	75.69

CNN = convolutional neural network; COPD = chronic obstructive pulmonary disease; ROC = receiver operating characteristic; AUC = area under curve; PR = precision-recall curve; GOLD = the Global Initiative for Chronic Obstructive Lung Disease; LAA = low attenuation area; FEV1 = forced expiratory volume in 1 s; FVC = forced vital capacity;

The bold fond is used to highlight the highest value for each column among different methods. Each row is a different method.

[a] We repeated the experiments on these methods and the results are reported on fivefold cross-validation over a dataset of 10 300 subjects.

[b] We reuse the results reported by Gonzalez et al.[20] The results are reported on a held-out set of 1000 subjects.

[c] COPD is diagnosed using model predicted FEV1/FVC > 0.7 and not as a binary classification.

[d] The GOLD-Stage is computed using decision tree classifier trained on predicted spirometry measurements.

[e] The ROC curve shows how the true positive (TP) vs false positive (FP) relationship changes as we vary the threshold of the positive class in our model. Higher AUC-ROC suggests better classification.

[f] Precision (TP/TP+FP) and recall (TP/TP+FN) quantifies the model's ability to identify instances from a positive class. High AUC-PR and recall indicate better identification of subject's with COPD.

Fig. 2). Next, we used the model-predicted FEV1/FVC ratio to diagnose COPD which achieved an AUC-ROC of 0.82. For the GOLD stage severity classification, our model achieved 65.4% and 89.1% exact and one-off accuracy's, respectively. Figure 2 shows the confusion matrix for the COPD-GOLD stage classification.

In Fig. 2(d), we visualize a random sample of the population by projecting the subject-level representations to two-dimensional (2D) space using a dimensionality reduction method called Uniform manifold approximation and projection (UMAP).[20] Uniform manifold approximation and projection provides a population-level view of the data while preserving local neighbor relations. Each dot in the scatter plot represents a patient, and its color denotes FEV1. The COPD disease severity increases with an increase in the temperature of the color. This plot confirms that, even in 2D embedding space, our model captures the disease; healthier subjects are visibly separable from severe subjects in the top left of the embedding space. As compared to other methods, the 2D embedding space for our methods visually looks much smoother and gradually transforms from healthy (yellow) to severe COPD subjects (blue). Also, there is much less overlap between the severity levels, and severe subjects are grouped in a visibly distinct cluster. Thus, the relative position of a subject in the 2D embedding space can monitor the COPD progression. It is also worth noting that the discrimination between severity groups is even higher as dimensionality increases. We used the 2D embedding for visualization purposes.

3.B. Visual emphysema score

Our model can identify subjects with different degrees of visual emphysema severity. The model correctly identified CLE visual emphysema score in 40.6% of the subjects in the COPDGene cohort and was within ± one score 74.8% of the time. Figure 3 compares the confusion matrices of our method and LAA-950 features. In staging paraseptal emphysema, the proposed model has an exact and on-off accuracy of 52.8% and 82.99%, respectively. Results are summarized in Table III, and the confusion matrix for paraseptal emphysema prediction is shown in Fig. 3. Application of the Hosmer–Lemeshow[36] test did not suggest evidence of poor calibration (P-value 0.079).

3.C. Acute eexacerbations

Our model achieved an AUC-ROC of 0.70 in identifying the subjects who reported experiencing at least one exacerbation before enrolling in the study. We compared our performance against the intensity-based LAA feature in Fig. 4 (results are summarized in Table IV).

We also evaluated the performance of our model in identifying the population who reported subsequent exacerbations at the time of the 5-year follow-up. Our model https://www.overleaf.com/project/5f20f578ff82f10001467e9d achieved an AUC-ROC of 0.68. Our experiments show that the previous exacerbation history, together with imaging features from our method performs better (AUC-ROC 0.73), in predicting future exacerbation events than using exacerbation history alone (AUC-ROC 0.67). A quantitative comparison between different methods is shown in Table IV. Figure 4 shows the ROC curve and the PR curve for binary classification. The P-value of the null hypothesis, using the Hosmer–Lemeshow test, is 0.08, suggesting no evidence of poor calibration.

3.D. mMRC dyspnea scale

Our proposed model was successful in classifying subjects in the COPDGene cohort based on their mMRC dyspnea scale with an accuracy of 43.5% and was within one score, 64.3% of the time (Table IV).

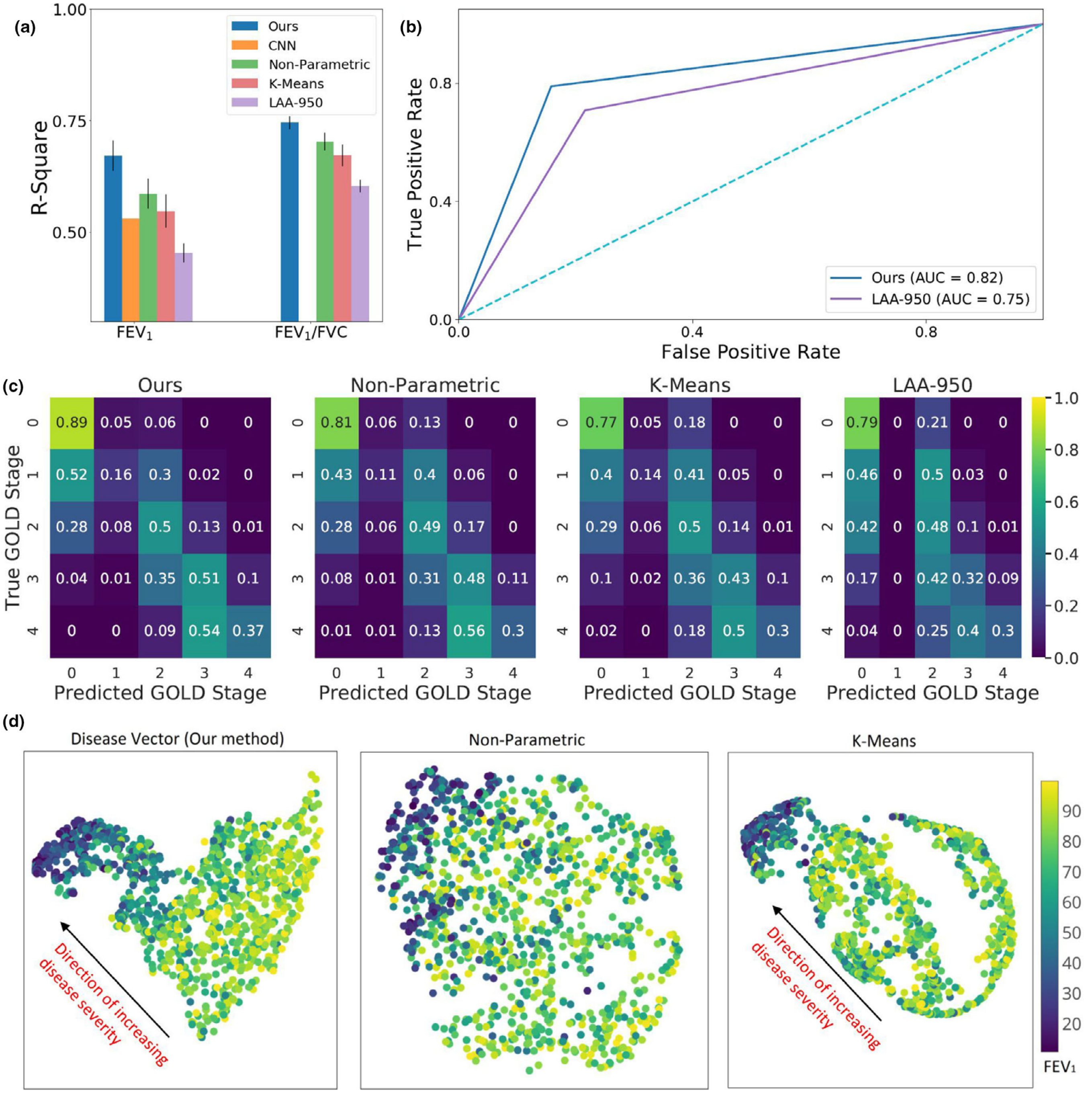

FIG. 2. Comparing different methods in predicting spirometry measurements, and COPD diagnosis and staging. (a) Bar graph comparing the r-square, coefficient of determination, for regression analysis of the forced expiratory volume in 1 s (FEV1) and FEV1/FVC, where FVC is the forced vital capacity. (b) Receiver operating characteristic (ROC) curve for prediction of COPD. The ROC curve shows how the true positive vs. false positive relationship changes as we vary the threshold of the positive class. Higher AUC-ROC suggests better classification. (c) Confusion matrix plot for staging subjects using the GOLD stage. Following the GOLD guidelines,[1] we used the model predicted FEV1 and FEV1/FVC ratio to diagnose and stage COPD. (d) Visualizing the population by projecting the patient-level representations to 2D space using a dimensionality reduction method called UMAP.[38] Each dot represents one subject colored by percentage predicted FEV1. The relative position of a subject can be used to monitor the progression. We use two dimensions for the sake of visualization; it is straightforward to use a higher dimension and improve patient characterization. Figure is best viewed in color.

3.E. Mortality

Our proposed method achieved a concordance of 0.61 in Cox regression[31] analysis compared to 0.56 for the BODE index and 0.53 for LAA-950 features (Table V). In testing the proportional hazard (PH) assumption of our model using scaled Schoenfeld residues, we achieved a $P > 0.3$ for all the covariates and a global P-value of 0.59 for the model. A significant p-value for this test provided no evidence for the violation of the PH assumption made by the Cox model. Next,

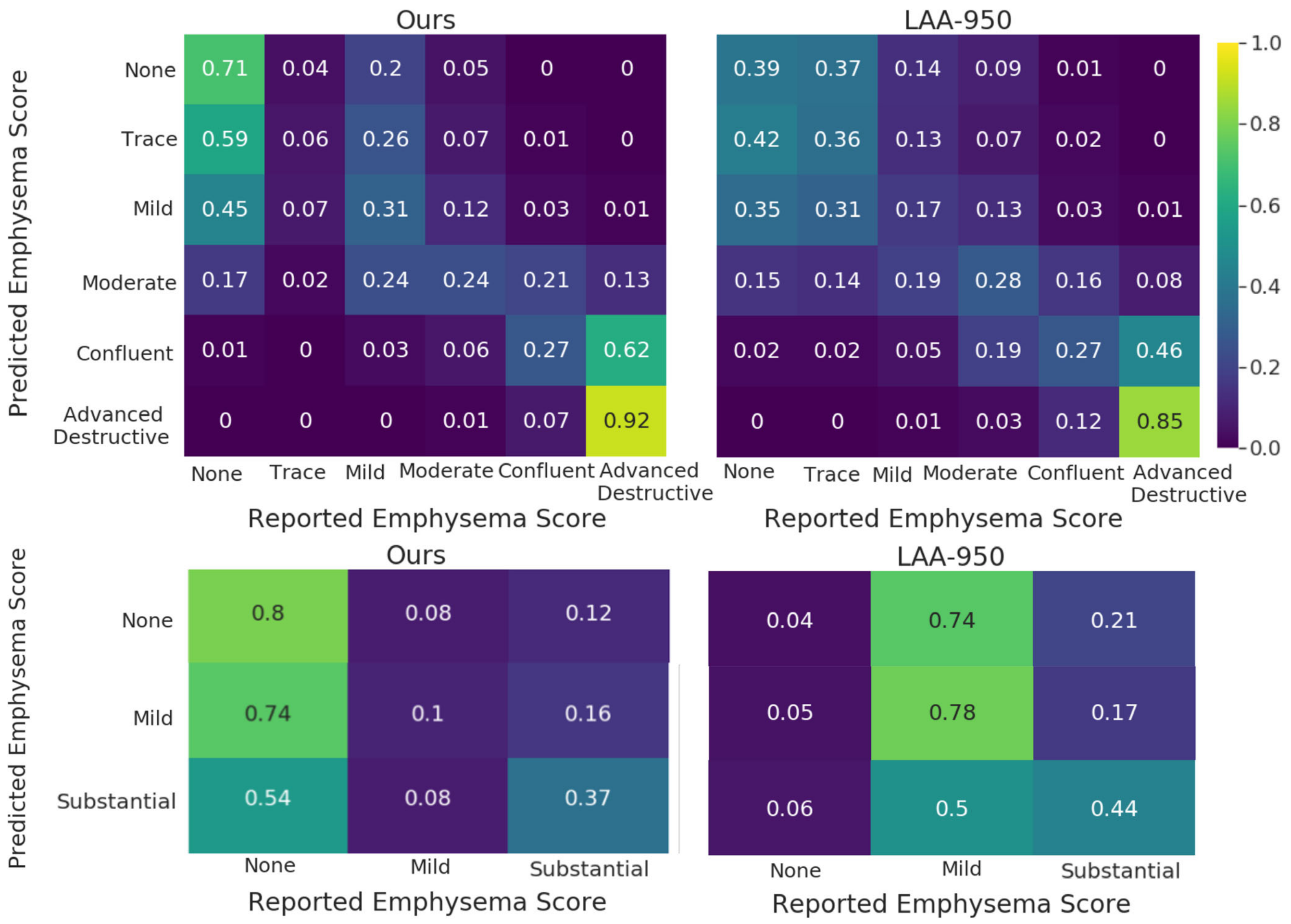

FIG. 3. Comparing our method against traditionally used computed tomography (CT) quantification measures (LAA-950) in stratifying the population-based on centrilobular and paraseptal emphysema severity score. Ours (direct) model is trained to predict spirometry measures and emphysema visual score together in a single loss function. The emphysema visual score is predicted in ordinal multi-class classification analysis. (a) Confusion matrix plot for grouping the COPD-Gene population-based on **centrilobular emphysema** and (b) **paraseptal emphysema**. Our proposed method performed better than LAA features and created a more significant separation between little and substantial emphysema.

we tested the global statistical significance of the Cox model using three alternative tests: the likelihood-ratio test, the Wald test, and the score log-rank statistic. We achieved a $P < 0.001$ in all three tests. Hence, we can reject the null hypothesis that all the coefficients are 0, with high confidence. Figure 5 shows the Kaplan Meier (KM) plots to visualize the subjects grouped by quantile of predicted probability of 5-yr survival. The KM plot for our method has a large separation between different quantile groups. Thus, our model can divide the population into distinct groups based on their survival risk.

4. DISCUSSION

Our proposed DL-based method demonstrates the ability to predict multiple aspects of COPD disease pattern, severity, and future events. It does so by extracting the most relevant information from volumetric HRCT images of the subject. Unlike previous DL methods that process a collection of 2D slices, our method works on the entire 3D inspiratory scan of the subject. DL enables us to go beyond standard radiographic features such as LAA and construct data-driven radiological features that are optimal for a specific task. Our results show that large cohorts such as COPDGene enable DL methods to learn meaningful patterns and converge to reliable predictions. Another advantage of our method lies in its generalizability and flexibility to incorporate different aspects of COPD. Using the same DL model and architecture, we were not only able to predict spirometric obstruction but were also successful in predicting all-cause mortality and future exacerbations, quantifying emphysema burden and disease pattern, and evaluating symptom scores.

In the direct approach, our model achieved high predictive strength by explicitly training to predict a target outcome. Our cross-validation experiments showed that the model was well calibrated and achieved consistent performance over all folds. While in the in-direct approach, the model was trained only once, to predict respiratory measurements, this model performed well in predicting COPD outcomes including, acute exacerbations, mortality, and mMRC.

TABLE III. Results classifying subjects based on their emphysema visual score.

	CLE[4]		Para-septal[4]	
Method[a]	% accuracy	% accuracy *one-off*	% accuracy	% accuracy *one-off*
Ours (direct)[b]	**40.61**	**74.68**	**52.82**	82.99
Ours (in-direct)[c]	36.30	61.33	46.87	75.97
Spirometry (FEV1)	33.52	63.96	44.64	72.77
LAA-950	31.89	77.74	33.32	**87.64**

LAA = low attenuation area; FEV1 = forced expiratory volume in 1 s; CLE = centrilobular emphysema;

The bold fond is used to highlight the highest value for each column among different methods. Each row is a different method.

[a]The results are reported on fivefold cross-validation over a dataset of 10 300 subjects.

[b]Ours (direct) model predicted spirometry measures and emphysema visual score together in a single loss function.

[c]Ours (in-direct) model predicted only spirometry measures as disease severity. The patient representations from this model are used in a separate multi-class classification analysis to predict the emphysema visual score.

Our predictions of spirometry measurements outperformed previously reported methods, including the previous DL method. Our method has a potential translational impact if it is utilized as a clinical screening tool, for example, when obtained during routine cancer screening, to identify subjects with a high likelihood of COPD for further assessment. Our visualization of the COPDGene population colored by the FEV1 value shows subjects with high FEV1 clustered together and a progression of disease severity from low to high [Fig. 2(d)]. This population-level analysis may be helpful in prospectively identifying unique clinical subgroups or in quantifying disease severity across research cohorts.

Our model's ability to predict future acute exacerbations has potential implications for health systems-level care given high costs associated with hospital admission for COPD exacerbations.[39] Currently, the strongest predictor of future exacerbations is the history of prior exacerbations.[40,42] Our DL model is complementary to the baseline exacerbation history in predicting future exacerbations. An automated DL approach offers scalability that could be used to identify high-risk patient pools for preemptive interventions such as medication optimization and pulmonary rehabilitation.

Currently, emphysema assessment requires manual scoring by trained radiologists. The manual process is very tedious and is prone to human error. Our model provides an objective way of identifying the visual score of emphysema from CT imaging which is more accurate than using CT quantification measures such as LAA features. Also, in a clinical setting, identifying a patient's emphysema severity by distinguishing between the extent of centrilobular and paraseptal emphysema can provide additional insights.

Notably, our model performed better than the BODE index and conventional emphysema quantification for mortality

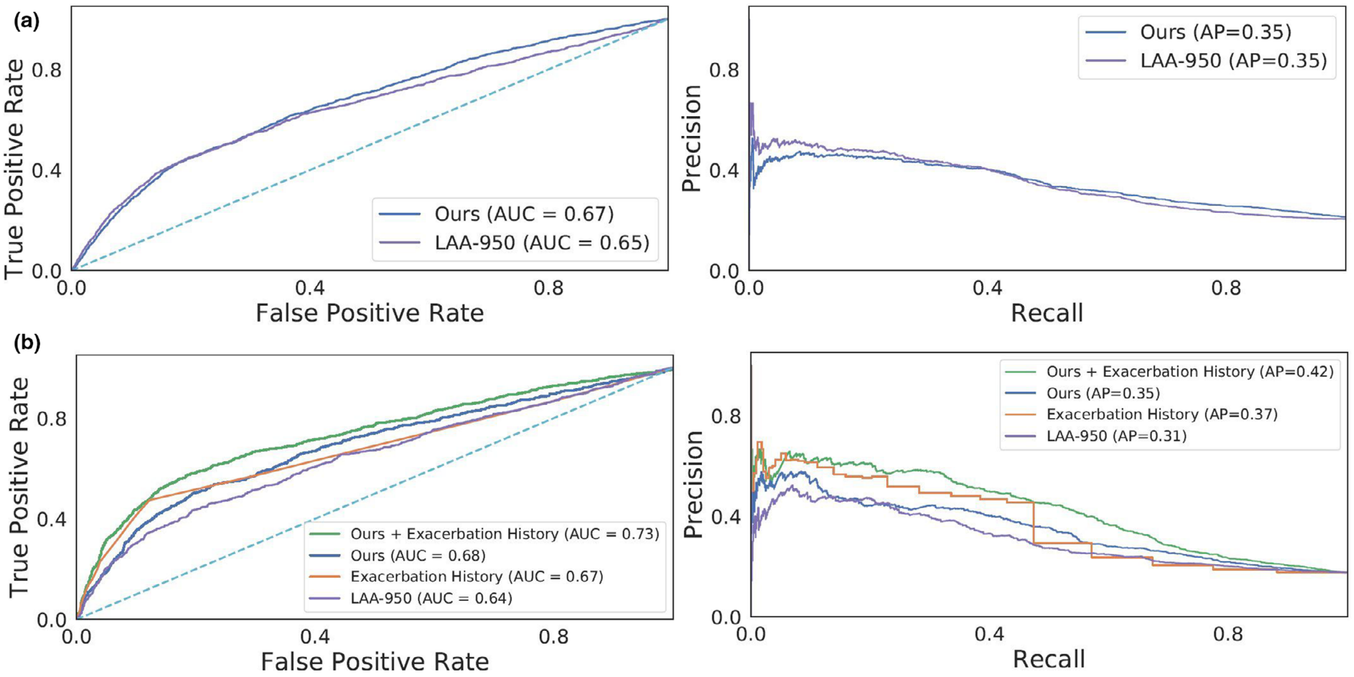

FIG. 4. Receiver operating characteristic (ROC) curve and precision-recall (PR) curve for identifying subjects with **A. exacerbation history** and **B. future exacerbation** as given in longitudinal follow up. The ROC curve shows how the true positive vs false positive relationship changes as we vary the threshold of the positive class. In the top row, the positive class represents those subjects in COPD Cohort who reported experiencing at least one exacerbation before enrolling in the study. In the bottom row, the positive class represents those subjects who reported experiencing at least one exacerbation at the 5-yr longitudinal follow up. Higher AUC-ROC number indicates better classification performance. Higher average precision (AP) in the PR curve means the better ability of the model in identifying subjects in a positive class. The plot shows that combining the history of past exacerbation with deep learning features from our model improves the prediction of future exacerbation. Figure is best viewed in color.

TABLE IV. Results for identifying subjects with exacerbation risk.

	Exacerbation history			
Method	ROC-AUC[e]	PR-AUC[f]	Recall	% accuracy
Ours (direct)[a,c]	0.68 ± 0.02	**0.38 ± 0.03**	0.27 ± 0.14	**76.93**
Ours (in-direct)[b,c]	**0.73 ± 0.01**	**0.43±0.03**	**0.59 ± 0.03**	74.75
CNN[20][d]	0.643	—	0.18	60.40
LAA-950	0.65 ± 0.01	0.35 ± 0.02	0.43 ± 0.02	73.78

	Future exacerbation in longitudinal follow-up			
	ROC-AUC	PR-AUC	Recall	% accuracy
Ours (direct)	0.65 ± 0.01	0.32 ± 0.02	0.43 ± 0.01	68.30
Ours (in-direct)	0.70 ± 0.02	0.35 ± 0.02	**0.57 ± 0.02**	73.87
LAA-950	0.64 ± 0.01	0.31 ± 0.02	0.43 ± 0.04	73.80
Exacerbation history	0.67 ± 0.02	0.37 ± 0.02	0.47 ± 0.04	80.60
Ours (in-direct) + exacerbation history	**0.73 ± 0.01**	**0.42 ± 0.02**	0.47 ± 0.04	**80.83**

	mMRC Dyspnea sscore[g]	
	% accuracy	% accuracy *one-off*
Ours (direct)	**46.40**	67.04
Ours (in-direct)	38.94	59.86
Spirometry (FEV1)	42.63	**69.07**
LAA-950	41.52	63.45

CNN = convolutional neural network; ROC = receiver operating characteristic; AUC = area under curve; PR = precision-recall curve; LAA = low attenuation area; The bold fond is used to highlight the highest value for each column among different methods. Each row is a different method.

[a] Ours (direct) model predicted spirometry measures and clinical outcomes of interest together in a single loss function.

[b] Ours (in-direct) model predicted only spirometry measures as disease severity. The generalized patient representations from this model are then used in a separate classification or regression analysis to predict other clinical outcomes.

[c] Results are reported on fivefold cross-validation over a dataset of 10 300 subjects.

[d] We reuse the results reported by Gonzalez et al.[20] The results are reported on a held-out set of 1000 subjects.

[e] The ROC curve shows how the true positive (TP) vs false positive (FP) relationship changes as we vary the threshold of the positive class in our model. Higher AUC-ROC suggests better classification.

[f] Precision (TP/TP + FP) and recall (TP/TP + FN) quantifies the model's ability to identify instances from a positive class. High AUC-PR and recall indicate better identification of subject's with COPD.

[g] Solved as ordinal multiclass classification. mMRC is a 5-category variable.

assessment, as seen in Table V. The BODE index is a commonly used mortality prediction tool for patients with COPD.[9] However, the requirement for a formal 6-minute walk distance test is a limitation in resource-poor settings and in patients with comorbidities that may interfere with walk performance. The ability to stratify patients based on mortality risk using imaging features derived from only a CT scan offers the opportunity to perform a large-scale population risk assessment and resource targeting.

The previous DL model,[20] performed similar analyses of mortality and exacerbation history prediction. Their CNN model analyzed only four slices from the chest CT and does not view the parenchyma as a volumetric object. In contrast our model, extract features from the entire volume, resulting in superior prediction for exacerbation and survival analysis. Gonzalez et al[20] reported their results on a held-out test set. In contrast, we performed a conservative evaluation of our method, and reported results on fivefold cross-validation. Our consistent results on the five folds rules out the possibility of over-fitting on COPDGene dataset as shown in our extended results in Supplementary Material.

There are limitations to this study, especially to the use of DL-based methods. Mostly, CNN-based DL models are opaque and hence, provides limited reasoning for a prediction. The adaptive weighting scheme component of our proposed model provides some insight into the model prediction, as saliency maps offer interpretability by identifying the areas assigned higher weights and thus viewed by the model as being more important in predicting disease outcomes. Through manual observation, we found that in severe COPD cases, the saliency map mostly focusses on the bullae area, although not always. It also picks up normal regions because the absence of the normal tissue suggests more destruction by the disease and hence, more severe emphysema. Despite this, further, improvement is required to make the DL methods more clinically interpretable.

A second challenge to our approach is how to use volumetric data from CT images more effectively. Our current approach represents a subject as an aggregation of 3D patches and does not account for spatial locations of the patches (regions). Such information is relevant as some emphysema visual subtypes have lung location bias; for

TABLE V. Results of Cox proportional-hazard (PH) model for survival analysis. The probability of death, learned from binary classification of mortality, is used as covariate in Cox regression.

Method	Hazard ratio[e]	Quantile *P*-value[f]	Concordance[g]	Global statistical significance[h] Max *P*-value (LR, Wald, log Rank)	PH-Assumption (Global *P*-value)[i]
Ours (direct)[a]	1.04 [CI: 0.09, 1.87]	<2e-16	0.590	*P* =< 2e-16	0.514
Ours (in-direct)[b]	1.54 [CI: 1.09, 2.17]	<2e-16	0.615	*P* =< 2e-16	0.598
CNN[20c]	2.69 [CI: 1.19, 6.05]	0.017	0.72	–	–
Spirometry (FEV1)	1.20 [CI: 0.94, 1.54]	6.91e-07	0.525	*P* = 4e-06	–
BODE index[8d]	1.68 [CI: 1.21, 2.31]	<2e-16	0.568	*P* =< 2e-16	0.462
LAA-950	1.13 [CI: 0.93,1.37]	6.35e-07	0.537	*P* = 4e-06	0.391

PH = proportional hazards; CNN = convolutional neural network; FEV1 = forced expiratory volume in 1 second; BODE = body mass index, airflow obstruction, dyspnea and exercise index; LAA = low attenuation area; CI = confidence interval.

All the models have age, gender, smoking pack-years, and center of enrollment as covariates.

The bold fond is used to highlight the highest value for each column among different methods. Each row is a different method.

[a] Ours (direct) model predicted spirometry measures and mortality together in a single loss function. Results are reported on fivefold cross-validation over a dataset of 10 300 subjects.

[b] Ours (in-direct) model predicted only spirometry measures as disease severity. The generalized patient representations from this model are then used in a separate binary classification analysis to predict mortality.

[c] We reuse the results reported by Gonzalez et al.[20] The results are reported on a held-out set of 1000 subjects.

[d] BODE index is the clinical index used to predict the mortality rate from COPD.[9]

[e] The Hazard ratio is the exponential coefficient (exp(β)) of the covariate. A covariate is positively associated with the event probability when the hazard ratio is above one and thus is negatively associated with the length of survival. We also report 95% confidence intervals for the hazard ratio.

[f] A significant *P*-value with > 1 hazard ratio indicates a strong relationship between the covariate and increased risk of death.

[g] The concordance shows the fraction of pairs, where the observations with higher survival time have a higher probability of survival predicted by the model. It is analog to the area under the ROC curve in classification analysis.

[h] The Global statistical significance of the model is tested using three alternative tests namely the likelihood-ratio (LR) test, the Wald test, and the score log-rank statistics. $P < 0.001$ indicates that the model fits significantly better than the null hypothesis. The null hypothesis states that all the betas (β) are 0.

[i] We used scaled Schoenfeld residuals to check the proportional hazards assumption. A non-significant *P*-value shows no evidence of violation of PH assumption by survival model.

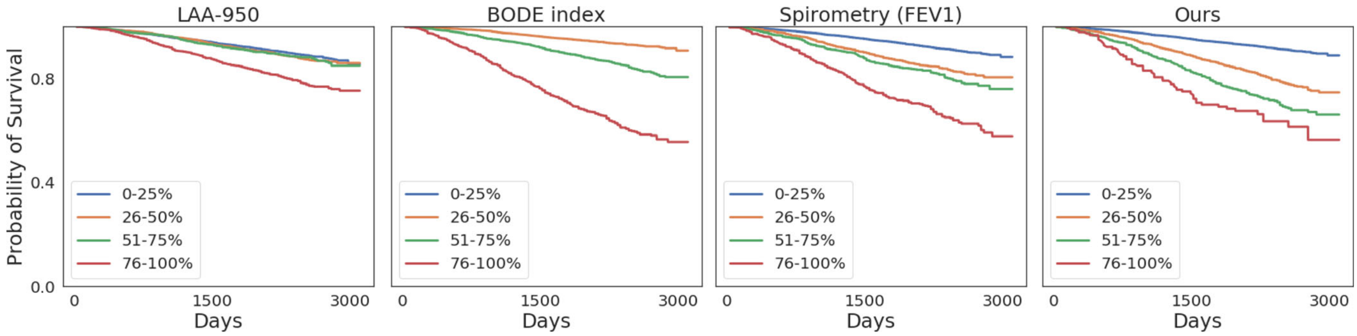

FIG. 5. Kaplan–Meier plot for visualizing the results of survival analysis. The plot is obtained by performing Cox regression analysis stratified on the quantile of predicted probability of mortality in binary classification. A good Kaplan–Meier plot has large separations between the groups. BODE index is the body mass index, airflow obstruction, dyspnea, and exercise index which is highly correlated with mortality.[9] Our model performed better than the conventional emphysema quantification, the BODE index, and spirometry measures for mortality assessment. Figure is best viewed in color.

example, the centrilobular emphysema is commonly located within the central portion of secondary pulmonary lobules.[43–45] Additionally, we extract features only from the parenchyma region of the chest CT scan. There is significant evidence that vasculature and bone and muscle structure are affected by the disease.[46,47] As a future direction, we plan to incorporate this additional information into our model.

5. CONCLUSIONS

This is the first study to use DL-based method to predicted various clinical outcomes associated with COPD like spirometric obstruction, emphysema severity, dyspnea extend, current and future exacerbation risk and mortality, using CI imaging alone. The results of our study conclude that DL-based method can provide a holistic view of disease severity and progression from a single set of CT images. Our model has potential applicability in both research and clinical practice. Further work toward developing interpretable DL models is essential for the development of standardized CT-based assessment of COPD.

[a)]Author to whom correspondence should be addressed. Electronic mail: sumedha.singla@pitt.edu.

REFERENCES

1. Vogelmeier CF, Criner G, Martinez FL, et al. Global strategy for the diagnosis, management, and prevention of chronic obstructive lung disease 2017 report. GOLD executive summary. *Am J Respir Crit Care Med*. 2017;195:557–582.
2. Coxson HO, Leipsic J, Parraga G, Sin DD. Using pulmonary imaging to move chronic obstructive pulmonary disease beyond FEV1. *Am J Respir Crit Care Med*. 2014;190:135–144.
3. O'Donnell DE, Laveneziana P. Physiology and consequences of lung hyperinflation in COPD. *Eur Respirat Rev*. 2006;15:61–67.
4. Grzela K, Litwiniuk M, Zagorska W, Grzela T. Airway remodeling in chronic obstructive pulmonary disease and asthma: the role of matrix metalloproteinase-9. *Arch Immunol Ther Experim*. 2016;64:47–55.
5. Bhalla AS, Das A, Naranje P, Irodi A, Raj V, Goyal A. Imaging protocols for CT chest: a recommendation. *Indian J Radiol Imaging*. 2019;29:236.
6. Washko GR, Hunninghake GM, Fernandez IE, et al. Lung volumes and emphysema in smokers with interstitial lung abnormalities. *N Engl J Med*. 2011;364:897–906.
7. Soler-Cataluña JJ, Martínez-García MA, Román Sánchez P, Salcedo E, Navarro M, Ochando R. Severe acute exacerbations and mortality in patients with chronic obstructive pulmonary disease. *Thorax*. 2005;60:925–931.
8. Celli BR, Cote CG, Marin JM, et al. The body-mass index, airflow obstruction, dyspnea, and exercise capacity index in chronic obstructive pulmonary disease. *N Engl J Med*. 2004;350:1005–1012.
9. Martinez FJ, Foster G, Curtis JL, et al. Predictors of mortality in patients with emphysema and severe airflow obstruction. *Am J Respir Crit Care Med*. 2006;173:1326–1334.
10. Lynch DA, Austin JHM, Hogg JC, et al. CT-definable subtypes of chronic obstructive pulmonary disease: a statement of the Fleischner society. *Radiology*. 2015;277:192–205.
11. Nishio M, Nakane K, Kubo T, et al. Automated prediction of emphysema visual score using homology-based quantification of low-attenuation lung region. *PLoS ONE*. 2017;12:e0178217.
12. Estépar RSJ, Kinney GL. Computed tomographic measures of pulmonary vascular morphology in smokers and their clinical implications. *Am J Respir Crit Care Med*. 2013;188:231–239.
13. Diaz AA, Valim C, Yamashiro T, et al. Airway count and emphysema assessed by chest CT imaging predicts clinical outcome in smokers. *Chest*. 2010;138:880–887.
14. Lynch DA, Moore CM, Wilson C, et al. CT-based visual classification of emphysema: association with mortality in the COPDGene study. *Radiology*. 2018;288:859–866.
15. Mohamed Hoesein FA, van Rikxoort E, van Ginneken B, et al. Computed tomography-quantified emphysema distribution is associated with lung function decline. *Eur Respir J*. 2012;40:844–850.
16. Müller NL, Staples CA, Miller RR, Abboud RT. Density mask: an objective method to quantitate emphysema using computed tomography. *Chest*. 1988;94:782–787.
17. Chen PY, Sharma Y, Zhang H, Yi J, Hsieh CJ. Ead: elastic-net attacks to deep neural networks via adversarial examples. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32, No. 1; 2018, April.
18. Sorensen L, Nielsen M, Lo P, Ashraf H, Pedersen JH, de Bruijne M. Texture-based analysis of COPD: a data-driven approach. *IEEE Trans Med Imaging*. 2012;31:70–78.
19. Schabdach J, Wells WM, Cho M, Batmanghelich KN. A likelihood free approach for characterizing heterogeneous diseases in large scale studies. In: *IPMI*. Vol. 10265. LNCS, 2017: 170–183.
20. Gonzalez G, Ash SY, Vegas-Sánchez-Ferrero G, et al. Disease staging and prognosis in smokers using deep learning in chest computed tomography. *Am J Respir Crit Care Med*. 2018;197:193–203.
21. Singla S, Gong M, Ravanbakhsh S, Sciurba F, Poczos B, Batmanghelich KN. Subject2Vec Generative-discriminative approach from a set of image patches to a vector, In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; 2018.
22. Regan EA, Hokanson JE, Murphy JR, et al. Genetic epidemiology of COPD (COPDGene) study design. *J COPD*. 2010;7:32–43.
23. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE*. 1998;86:2278–2324.
24. San Jose Estepar R, Ross JC, Harmouche R, Onieva J, Diaz AA, Washko GR. Chest imaging platform: an open-source library and workstation for quantitative chest imaging. In: *C66. Lung Imaging II: New Probes and Emerging Technologies*. Vancouver: American Thoracic Society; 2015:A4975–A4975. https://www.atsjournals.org/doi/abs/10.1164/ajrccm-conference.2015.191.1_MeetingAbstracts.A4975
25. Zaheer M, Kottur S, Ravanbakhsh S, Poczos B, Salakhutdinov R, Smola A. Deep sets. In: *Advances in neural information processing systems*; 2017:3391–3401.
26. Masci J, Meier U, Cireşan D, Schmidhuber J. Stacked convolutional auto-encoders for hierarchical feature extraction. In: *International Conference on Artificial Neural Networks*. Berlin, Heidelberg: Springer; 2011, June:52–59.
27. Clevert DA, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289; 2015.
28. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980; 2014.
29. Ash SY, Harmouche R, Vallejo DLL, et al. Densitometric and local histogram based analysis of computed tomography images in patients with idiopathic pulmonary fibrosis. *Respir Res*. 2017;18:45.
30. Qureshi H, Sharafkhaneh A, Hanania NA. Chronic obstructive pulmonary disease exacerbations: latest evidence and clinical implications. *Ther Adv Chron Dis*. 2014;5:212–227.
31. Perez T, Burgel PR, Paillasseur JL, INITIATIVES BPCO Scientific Committee, et al. Modified medical research council scale vs baseline dyspnea index to evaluate dyspnea in chronic obstructive pulmonary disease. *Int J Chron Obstruct Pulmon Dis*. 2015;10:1663.
32. Lin DY. Cox regression analysis of multivariate failure time data: the marginal approach. *Stat Med*. 1994;13:2233–2247.
33. Schoenfeld D. Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika*. 1980;67:145.
34. Davidson-Pilon C, Kalderstam J, Jacobson N, et al. CamDavidsonPilon/lifelines: v0.25.7 (Version v0.25.7). Zenodo; 2020. https://doi.org/10.5281/zenodo.4313838
35. Therneau TM, Grambsch PM. The Cox model. In: *Modeling Survival Data: Extending the Cox Model*. New York: Springer; 2000:39–77.
36. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*. 2015;10:e0118432.
37. Lemeshow S. Hosmer DW. A review of goodness of fit statistics for use in the development of logistic regression models. *Am J Epidemiol*. 1982;115:92–106.
38. Frank E, Hall M. A simple approach to ordinal classification. In *European Conference on Machine Learning*. Berlin, Heidelberg: Springer; 2001, September:145–156.

39. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv: 1802.03426.
40. Guarascio AJ, Ray SM, Finch CK, Self TH. The clinical and economic burden of chronic obstructive pulmonary disease in the USA. *Clin Econ Outcomes Res: CEOR*. 2013;5:235–45.
41. Rothnie KJ, Müllerová H, Smeeth L, Quint JK. Natural history of chronic obstructive pulmonary disease exacerbations in a general practice-based population with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*. 2018;98:464–471.
42. Rao BD, Kreutz-Delgado K. An affine scaling methodology for best basis selection. *IEEE Trans Signal Process*. 1999;47:187–200.
43. Hurst JR, Vestbo J, Anzueto A, et al. Susceptibility to exacerbation in chronic obstructive pulmonary disease. *N Engl J Med*. 2010;363: 1128–1138.
44. Anderson AE, Hernandez JA, Holmes WL, Foraker AG. Pulmonary emphysema. *Arch Environ Health*. 1966;12:569–577.
45. Snider GL. Pulmonary tuberculosis and centrilobular emphysema: the “Upright Theory” of apical localization. *Arch Intern Med*. 1963; 111:762–771.
46. Smith BM, Austin JH, Newell JD, et al. Pulmonary emphysema subtypes on computed tomography: the MESA COPD study. *Am J Med*. 2014;127:7–94.
47. Gea J, Pascual S, Casadevall C, Orozco-Levi M, Barreiro E. Muscle dysfunction in chronic obstructive pulmonary disease: update on causes and biological findings. *J Thorac Dis*. 2015;7:E418.
48. Cielen N, Maes K, Gayan-Ramirez G. Musculoskeletal disorders in chronic obstructive pulmonary disease; 2014.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1. (a) Spectral properties of patch-level features for different values of λ_1. (b) The trade-off between rank of the latent space (red, y-axis on left) and the predictive power (blue, y-axis on right) for different values of λ_1. Left represents fully discriminative ($\lambda_1 = 0$) and right represents fully generative models ($\lambda_1 \rightarrow \infty$).

Data S1. 1.A: Architectural Details, 1.B Ablation Study 1.C Extended Results.