An unsupervised anomaly detection model to identify emphysema in low-dose computed tomography

Yeshaswini Nagaraj, *Member, IEEE*, Ludo Cornelissen, Jiali Cai, Mienieke Rook, Hendrik Joost Wisselink, Raymond Veldhuis, Matthijs Oudkerk, Rozemarijn Vliegthart, and Peter van Ooijen, *Senior Member, IEEE*

Abstract—Challenges such as class imbalance, time-intensive visual scoring, and limited amounts of labeled data are often encountered while accessing lung cancer screening low-dose computed tomography (LDCT) data for automated emphysema detection. To tackle these issues, we propose a generative adversarial network (GAN) based on unsupervised anomaly detection (UAD) to identify emphysema in LDCT. And to initiate disease specific feature learning, we introduce data augmentation based on minimum intensity projection (minIP) for the adversarial framework. The generator of the UAD model resembles an image-latent-image configuration consisting of fully convolutional encoder-decoder architecture with skip connections. The discriminator is an encoder, which acts as a feature extractor and a classifier. We tested the robustness of the minIP based UAD model by adding anomalies to the training set. Furthermore to illustrate the usefulness of minIP and to serve as a comparison methodology for emphysema detection in LDCT, the proposed model was compared to the 2.5D transfer learning model. Visualization techniques like post-processed residual images and Grad-cam maps were used to explain the UAD model inference. Our proposed minIP based UAD model showed an area under the curve of 0.91 ± 0.02 for emphysema detection in LDCT. The UAD model was robust to inclusion of anomalies up to 3%. The augmentation improved the sensitivity of the transfer learning model by 2% however the UAD model outperformed the transfer learning model. The UAD model with minIP is a potential computer-aided tool for early detection of emphysema in lung cancer screening data.

Index Terms—Anomaly detection, Emphysema, Generative adversarial networks, Disease-specific augmentation

Y.Nagaraj, L.Cornelissen and P. van Ooijen are with University of Groningen, University Medical Center Groningen, Department of Radiation oncology and Machine learning lab, DASH, Groningen, Groningen, The Netherlands

J. Cai is with University of Groningen, University Medical Center Groningen, Department of Epidemiology, Groningen, The Netherlands

H.J. Wisselink, M. Rook and R. Vliegthart are with University of Groningen, University Medical Center Groningen, Department of Radiology, Groningen, The Netherlands

M. Rook is with Martini Hospital Groningen, Department of Radiology, Groningen, The Netherlands

R. Veldhuis is with the University of Twente, Faculty of Electrical Engineering, Mathematics & Computer Science (EWI), Data management & Biometrics (DMB), Enschede, The Netherlands

M. Oudkerk is with University of Groningen, Faculty of Medical Sciences, Groningen, The Netherlands

M. Oudkerk is with Institute for Diagnostic Accuracy Research B.V., Groningen, The Netherlands

I. INTRODUCTION

FOR early detection of lung cancer, hospitals in Europe and the USA have adopted lung cancer screening trials using low-dose computed tomography (LDCT) [1, 2]. Although lung cancer screening trials have been credited with significant reduction in lung cancer-related mortality, early detection of the other "Big 3" (B3) lung diseases – namely, chronic obstructive pulmonary disease (COPD) and coronary artery disease – using lung cancer screening LDCT (LCS-LDCT) can potentially help to reduce mortality in the overall population [3]. Two main components characterize COPD: airway wall thickening, or bronchitis, and lung parenchymal destruction, or emphysema [4]. Screening studies have shown increased lung cancer detection rates in cases of emphysema [5]. However, accessing large amounts of LCS-LDCT data on emphysema requires extensive visual assessment. The complexity of the image interpretation due to noise and streak artifacts, combined with the time demands on radiologists, makes emphysema detection labor-intensive and time-consuming. Therefore, for detecting an asymptomatic disease like emphysema, where destruction of the lung tissue is irreversible, the available LCS-LDCT data could be leveraged by applying data-driven approaches to ease radiologists' burden for timely implementation of therapeutic measures.

Although deep learning models are well known for automating time-intensive tasks, choosing the right model for a target dataset before deploying it is vital. We therefore considered data-related properties like imbalanced data and lack of labels to select the right model for efficient detection of emphysema in LDCT. Imbalanced data: Discrimination of emphysema in trials screening for lung diseases is made difficult by an enormous class imbalance problem since the percentage of scans containing emphysema is small when compared to non-emphysematous scans [6]. Several studies have discussed anomaly detection algorithms to exploit class imbalance as an advantage and demonstrated their prospects in medical applications [7, 8]. A common hypothesis for anomaly detection models is that the deep learning network is trained on normal samples to learn the distribution of normality. When the network is tested on an unseen sample that departs from the norm, the network detects it as an anomaly by providing a high anomaly score. Therefore, we considered that, given

the class-imbalanced LDCT dataset, all the scans except the ones containing emphysema should be used as the major class to train the anomaly detection model to discriminate emphysema scans as an anomaly. Lack of labels: Emphysema annotation is not the primary focus of lung cancer screening trials, and visual scoring is time-intensive and expensive as LDCT has low sensitivity compared to HRCT. Most of the screening participants are healthy, or have mild emphysema or only traces of it, which makes annotating emphysema more challenging [9]. Therefore, models with unsupervised learning could be favoured over semi-supervised or supervised approaches to bypass emphysema annotation in LSC-LDCT. Despite the fact that unsupervised training does not involve labels, previously published unsupervised anomaly detection (UAD) models require clean or anomaly-free training datasets [7, 8, 10]. In a clinical setup, one may not have access to a clean dataset without radiologist or expert supervision. Therefore, in this study, apart from using a clean training dataset, we also tested our model's robustness by training the model with a training dataset contaminated with anomalies.

Recently, GANs [11] have been successfully applied to model complex, high-dimensional problems [12]. GANs made their appearance in medical imaging once their capability to learn semantic image content and efficient image-to-image translation became clear [13]. The added advantage of model training via the adversarial scheme is that it neither requires manual designing of higher-order losses (constraints or cost functions), nor does it introduce additional architectural complexity to the model [14]. This motivated us to employ a GAN to perform anomaly detection. Additionally, the images generated by the network can be qualitatively validated to build a reliable one-class model. We identified minimal differences between non-emphysematous and emphysematous scans, and as a method to emphasize these differences, we introduce the use of emphysema-specific data augmentation based on minimum intensity projection (minIP) for the adversarial framework. This domain knowledge is used by radiologists to enhance the hypo-attenuation regions of CT for better differentiation between non-emphysematous and emphysematous regions [15, 16].

The key contributions of this work are summarized as follows:

- An LDCT dataset was leveraged for automatic emphysema detection using the generative adversarial network-based unsupervised anomaly detection (UAD) model.
- For efficient discrimination between non-emphysematous and emphysematous scans, we introduce the minimum intensity projection (minIP) based data augmentation to the input scans.
- To mimic the case of truly unsupervised learning and to evaluate the model robustness, model performance was calculated by adding contamination during training.
- Qualitative assessment of the model was performed using visual Turing test and, the visualisation of discriminative features via Grad-cam and residual images was evaluated.
- To further understand the effect of minIP, UAD model was compared with a 2.5D transfer learning model.

The rest of the paper is organized as follows: In section II a brief overview of existing automatic detection models for emphysema and GANs in medical imaging is described. Section III describes the materials and methods like data characteristics, disease-specific augmentation, unsupervised anomaly detection, anomaly visualization and transfer learning model used for the experiments. In section IV, we present the experimental design and results, and the section V provides the discussion and conclusion to the work.

II. RELATED WORKS

1) *Overview of automatic detection of emphysema*: In clinical routine, emphysema is assessed using combination of lung function tests and visual confirmation on high resolution CT (HRCT) [17]. Recently, automatic detection of emphysema has shown potential in aiding the clinical routine [18]. Computer aided emphysema detection has been performed either by using traditional machine learning or deep-learning techniques. These models evaluate the local intensity distribution or use texture-based information to discern emphysema in CT [19, 20]. Deep learning models, such as 3D convolutional neural networks (CNNs), deep-CNNs with long short-term memory and transfer learning models like 3D ResNet have been found to detect emphysema with acceptable performance [21, 22, 23]. However, all these approaches primarily make use of HRCT and there is a dearth of studies on low-dose CT for automatic detection of emphysema. Lung cancer and COPD screening trials around the world are employing LDCT due to less radiation exposure compared to HRCT [24]. By developing automatic emphysema detection model using LDCT, LDCT can be leveraged to improve lung cancer risk prediction while detecting individuals with COPD.

2) *GANs in medical imaging - detection and detection*: The capability of GANs has allowed anomaly detection in highly class-imbalanced and unlabeled medical data facilitating an unsupervised [14, 25, 26] or a semi-supervised [27] anomaly detection model over supervised anomaly detection. Besides these capabilities, GAN architecture can be tweaked for performance gain based on the task at hand. One such example is unsupervised anomaly detection by [8] on optical tomography images, where the Deep convolutional GAN (DCGAN) discriminator architecture was replaced with a Wasserstein GAN-gradient penalty (WGAN-GP) to improve the performance of their previous architecture [14]. Additionally, to improve the image to latent space mapping of their GAN, they replaced the slow iterative restoration method with a single forward pass through the network [28]. Furthermore, Udrea et al. used encoder and decoder architecture as a generator in conditional-GANs (c-GANs) to enhance the performance of skin lesion detection via adversarial training [25].

GANs are challenging to train as their training stability depends on two networks, the generator and the discriminator, to reach equilibrium and hence is dependent on the type of loss function levied [29]. Goodfellow et al. formulated the adversarial loss based on "Jensen-Shannon distance" to measure the difference between generated and real data distribution, however, this showed convergence issues and caused training

instability [30][31]. To overcome this issue, a loss function was formulated by measuring the Wasserstein distances between the probability distribution of the input and the reconstructed image [31] and this was further improved by [32] for better latent representation. Recently, this was used in medical imaging by [26] and [8] for different classification and detection applications. An alternative approach to stabilize GAN training is to use fully convolutional layers and batch normalization, which was introduced by [33, 34]. We incorporated the fully convolutional layer approach with batch normalization in each layer throughout the network to ensure stability during training. Additionally, adversarial autoencoders (AAEs) are inherently capable of handling latent space [35]. Apart from using AAEs, we adopted the loss function proposed by [34], where the loss is defined as a weighted sum of adversarial, latent, and contextual loss. These losses ensure high fidelity, capturing the probability distribution of input image representation both in latent and image space. This combination of loss, where pixel-based l_1 and l_2 loss (contextual and latent loss) combined with adversarial loss, is found to improve the generator's efficiency in generating context-relevant images while maintaining training stability [12].

Fig. 1. The flowchart indicates the inclusion criteria and the data split for the training and testing. All the scans used for current study are the baseline or earliest scan available for each participant.

III. MATERIALS AND METHODS

A. Data description

The dataset was acquired from a general population-based study called "ImaLife", which comprised of participants aged 45 years and above from the northern part of the Netherlands. The primary goal of the ImaLife study was to establish CT reference values of quantitative imaging markers for the

"Big 3" lung diseases. Study participants of this study underwent a low-dose chest CT examination on a third-generation dual-source CT (SOMATOM Force, Siemens, Germany). The dataset consisted of scans (a minimum of 300 slices per scan) reconstructed in the axial plane with filtered back projection using the Br40 kernel (qualitative-smooth). The acquisition protocol for the dataset has been previously published [36]. The detailed consort of dataset used in the study is provided in the Fig.1.

Radiologists randomly selected two hundred and seventy five scans from the dataset, which were annotated as non-emphysematous (190) or emphysematous (85) based on the Fleischner criteria [37]. Abnormality included different levels of severity. A participant scan was marked as abnormal if one or more slices were marked as emphysematous, and marked as normal otherwise. The random selection and visual scoring of the ImaLife subcohort was performed by two radiologists (>5 years of radiological experience) and a technical physician (<2 years of radiological experience). From the independent sessions of annotation no overlap scans were reported.

B. Disease-specific data augmentation

The acquired ImaLife subcohort was considered for the experiments with and without disease-specific augmentation. We used minimum intensity projection (minIP) as the disease-specific augmentation technique to differentiate the low attenuation regions in LDCT. The minIP algorithm projects the voxel with the lowest attenuation onto every window throughout the volume, as illustrated in Fig. 2. As a result when the minIP algorithm is applied, the darkest voxels are visualized representing the minimum Hounsfield value across the slab. Therefore, in LDCT, the grayscale resolution or the contrast differences between with and without emphysema are prominently represented [16, 38]. For minIP processing on slice level we used an in-house Python algorithm followed by other standard augmentation techniques described in Section IV. An example of minIP processing on the normal and the emphysematous scans is shown in Fig. 3.

C. Unsupervised anomaly detection

The unsupervised anomaly detection (UAD) model is composed of generator and discriminator architecture. The generator (G) consists of adversarial autoencoders (AAEs) as encoder (V_e) and decoder (V_d) networks. The minIP processed LDCT scans are fed to the V_e network, which consists of eight fully convolutional downsampling layers with a kernel size of 4×4 and stride 2, where each layer is followed by batch normalization and Leaky ReLU activation functions. The downsampling layers of V_e are sequentially connected over short-ranged connections to output a low-dimensional feature representation of the input scans in the latent space (the bottleneck features, represented as a dark blue bar in the architecture diagram shown in Fig. 4). The bottleneck features are then converted into high-dimensional image with the V_d network's eight upsampling transposed convolutional layers. The upsampling layers are connected to each other with short-ranged connections followed by batch normalization and

Fig. 2. Illustration of minimum intensity projection on an axial chest LDCT scan in the z-direction. The processed scan contains the darkest voxels of the 11 slices, which represents low attenuation regions in the lung parenchyma. The pure black and pure white are the darkest and lightest shade of gray in the imaging scale.

Fig. 3. Illustration of scans before and after minIP processing. The emphysema regions are indicated using red bounding boxes by a radiologist. It can be observed that application of minIP improves the emphysema representation and suppresses the high intensity regions in parenchyma like blood vessels which is not important for emphysema detection (the image contrast has been adjusted for better visualisation of lung parenchyma and for interpretation of the references to colour in the main text, the reader is referred to the web version of this article.)

ReLU activation functions. For efficient recovery of high-dimensional data and improved gradient flow, we connected the upsampling layers of the encoder and the downsampling layers of decoder with long-ranged skip-connections. The outer layers of V_e and V_d represent shallow-level features, while more high-level features are recognized by deeper layers. In our architecture, we used dropout sampling in the skip-connections connecting the deeper layers while retaining all the neurons in the shallow layers. The generator generated high-dimensional image, and the corresponding input scan are taken by the discriminator which consists of 8 down-sampling convolution layers followed by batch normalization and Leaky ReLU functions, to provide feedback to the generator for minimizing the model loss. Thus the G learns to generate $\bar{x}$ from the input x of dataset S_{train} . This increases the difficulty for the D to differentiate between input and reconstructed images, initiating a backward learning mechanism to classify

input and generated images. The D is not just a classifier: it is also a feature extractor that can extract features within the latent space by forward inference [33]. The knowledge gained from backward learning and forward inference – that is, adversarial loss (Ψ_A) and latent loss (Ψ_L), respectively – are used to penalize the G , forcing it to learn the normality of the given dataset. The combination of latent (feature representation) and adversarial loss helps to minimize instability in training [29]. For latent loss (Ψ_L), the latent information from the penultimate convolutional layer of the D ($f(\cdot)$) was considered, where f represents a layer of the D used to extract a feature representation of the $f(x)$ and $f(\bar{x})$ [30]. Additionally, as indicated by [12], for better visualization of images, the pixel-wise Manhattan (l_1) norm was imposed between input and generated images as contextual loss (Ψ_C) [34]. The empirical loss is a weighted sum of all three loss functions.

$$\Psi = \lambda_A \Psi_A + \lambda_C \Psi_C + \lambda_L \Psi_L \quad (1)$$

where $\lambda_A, \lambda_C, \lambda_L$ are the weights applied for adversarial training and control how much each loss impacts the training. The optimum weights are defined after experimentation. The UAD model architecture is shown in Fig. 4.

Detection of anomalies: In the second phase, the trained model provides an anomaly score $A\{\cdot\}$ at the scan level for a given input sample x' by concatenating the scores from the sequence of axial slices for the final participant score. The anomaly score is low, or ideally zero, for $x' \in S_{test}$ if x is closer to the normal distribution learned from S_{train} . Otherwise, the model is expected to put out a higher anomaly score. We use the anomaly equation as defined by [14, 39]. Anomaly score is based on the reconstruction error, which is calculated using the contextual similarity $\dot{C}$ and latent space similarity $\dot{L}$ between input and generated images. For a given test image x' anomaly score is defined as:

$$A(x') = \lambda \dot{C} + (1 - \lambda) \dot{L} \quad (2)$$

where $A(x')$ ranges from 0 to 1, and λ is the loss weighing factor.

Visualization of anomalies: In the generated anomalous (emphysematous) images ($\bar{x}'$), the model, which knows the normal anatomy fails to reconstruct the anomaly structures. When this incorrect recovery of anomalous scans is subtracted

Fig. 4. The proposed unsupervised anomaly detection GAN architecture for emphysema detection in LDCT. The architecture works in an adversarial scheme which includes a generator block and a discriminator block. (For interpretation of the references to colour in the main text, the reader is referred to the web version of this article.)

from the input images (x'), the residual images are obtained ($x' - \bar{x}'$) [28]. The post-processed residual (PPR) images were obtained by binarizing the residual images ($x' - \bar{x}'$) and performing morphological operations [40].

Furthermore, to understand the model better, we visualized the discriminative image features of gradient information from the last convolutional layer of the discriminator [41]. This was realized using the following equation:

$$H^c_{grad} = ReLU(\sum_k w^c_k A^k) \quad (3)$$

where H^c_{grad} represents the discriminative feature map of the image and w^c_k is the weight in the last convolutional layer corresponding to the k^{th} feature map activation A^k belonging to class c . The grad-cam visualisation is drawn from the penultimate layer $f(\cdot)$ of the discriminator, which produces k feature maps.

D. Transfer learning approach

To further understand the effects of disease-specific augmentation, we compared our proposed model to a traditional transfer learning model in – with and without minIP settings. For that we choose a 2.5D Residual Neural Network (ResNet) model [42]. The network has 49 convolutional layers and only one fully connected layer connected through skip-connections followed by batch normalization and a ReLU activation function.

E. Evaluation metrics

In our experiments, to investigate the diagnostic performance of UAD and transfer learning model for emphysema detection, area under receiver operation curve (AUC) was used. In addition, standard performance evaluation metrics like F1 measure, sensitivity, specificity and accuracy were examined to compare the detection performance of with and without minIP scenario. The 95% confidence interval (CI) for the metrics was calculated using non-parametric bootstrapping with 1000 iterations. Class separation histogram plots and percentage overlap were calculated as suggested by [26] between normals and abnormals for with and without-minIP processing scenarios to further understand the effect of minIP processing on LDCT for emphysema detection. Finally, to check the robustness of the UAD model in case of contamination or underestimation of emphysema labels, we evaluated the performance of the minIP based UAD model using AUC for each percentage of intentional inclusion of contamination. We incorporated visual Turing test for quantitative assessment of UAD generated images, where false recognition percentage (FRP), and true recognition percentage (TRP) were calculated. The FRP represented the number of generated images recognized as generated by the radiologists and, the TRP represented the number of real images rated as real.

IV. EXPERIMENTS AND RESULTS

A. Data split and model implementation

The dataset under consideration was split into training (S_{train}) and testing (S_{test}) datasets. S_{train} contained only

non-emphysematous scans representing the normality. S_{test} contained non-emphysematous scans and scans with emphysema. To train the proposed UAD model, 150 scans (minimum of 300 slices per scan) from dataset S_{train} were used. Along with 40 remaining non-emphysematous scans, an equal number of emphysema scans were used for testing (S_{test}).

For the transfer learning model 90 scans (45 normals [45/150] and additional 45 abnormal scans) were used for training. The test split for the transfer learning model was kept the same as for the unsupervised model (i.e., 80 scans) and 2 fold-cross validation was carried out to evaluate the transfer learning approach. The additional scans of 45 emphysema were further used for contamination evaluation of the UAD model, by keeping the testing environment same for all the model settings.

Standard data augmentation: The entire dataset was normalized using the zero mean and unit variance methods. Furthermore, to ensure diversification of the training dataset and to avoid overfitting of the discriminator, standard data augmentation (ordered data rotation with 0° , 90° , 180° and 270° settings with uniform probability) [43], along with horizontal-vertical flip, contrast adjustment, Gaussian filters, and elastic deformation with $\sigma = 3$ was applied. These augmentation were performed irrespective of whether the training datasets are processed with disease-specific augmentation for both UAD and transfer learning model.

The model was implemented on PyTorch (version 5.1, Python 3.7.1, CUDA 10.1) and NVIDIA Titan XP GPU with 16GB memory. The training objective was optimized via the Adam optimizer with a batch size of 8. The learning rate used in the study was initially set to 2×10^{-2} and 2×10^{-6} with a 0.01 decreasing factor for both adversarial and transfer learning models. For adversarial training, lambda decay was used as a regularization term, and the momentum β_1 was set to 0.5 after several experiments using values ranging from 0.1 to 0.9 for stable training. The initial weighting parameters λ_A , λ_C , and λ_L were set to 1, 50 and 1, respectively. The λ in equation 2 was varied (0.1, 0.3, 0.5, 0.9) to find the right λ for our detection task. We relied on the highest AUC to choose the right λ . The number of training iterations was fixed to 300 epochs. The early stopping criterion was triggered if the model performance on the test set did not improve after 10 epochs. For both settings, with and without minIP, the networks showed no signs of mode collapse as the training accuracy increased with a decrease in the reconstruction loss.

B. UAD model evaluation

Model performance: The minIP-UAD model detected emphysema with the highest accuracy of 0.88 compared to the model evaluated without-minIP (accuracy of 0.80). The minIP processing boosted the AUC by 12% when compared to the without-minIP model confirming the advantage of disease specific augmentation. The evaluation metrics for with and without scenarios are shown in Table I. The overlap percentage between normal and abnormal distributions, showed that with minIP the percentage overlap was smaller (13%) when compared to without-minIP (26%), indicating a better class

separation after minIP processing (Table II). The AUC along with the stacked histogram plot are illustrated side by side in Fig. 5.

Model vulnerability for contamination: In our case, this means to evaluate the model performance in the presence of emphysema scans in training datasets. We contaminated the training dataset of the minIP based UAD model with 1%, 3% and 5% anomalies. Consequently a performance decrease in area under the curve to 0.89 ± 0.06 , 0.84 ± 0.07 and 0.82 ± 0.06 was observed respectively. Therefore, in our case the intentional inclusion of anomalies greater than 3% drastically decreased the performance by 8%.

Visual Turing test: We incorporated the visual Turing test to validate the quality of generated images and to understand how well the model learns the normal class distribution (S_{train}). For the visual Turing test, we first acquired the radiologists' opinion on 60 images (30 input CTs/real and 30 generated CTs) of size 256×256 . We then had two domain experts rate the randomized images as real or generated independently. The FRP for the UAD-generated images was 41.7% for expert 1 and 12.5% for expert 2, which means 58.3% and 87.5% of generated images were rated as real. This ensured that the UAD model is able to generate input images convincingly by capturing the S_{train} image distribution. The TRP was 61.2% and 72.2% for experts 1 and 2, respectively indicating the high sensitivity in recognising real as real by our experts.

Visualization of anomalies: The PPR images correctly represented the emphysema regions, but also sometimes showed the low-attenuation regions of the lung such as air trappings in trachea and bronchus segments. The examples of PPR images for the selected anomaly images are shown in the second row of Fig. 6. Grad-cam feature maps are visualized as heat maps in Fig. 6, and the most discriminative feature areas are colored red. The UAD grad-cam based on an average of the 11 slices (minIP) were consistent with the radiologist's ground truth (red bounding box).

C. Comparison of the unsupervised approach with transfer learning

The ResNet model showed an increased AUC of 2% with minIP when compared to without minIP. Table I shows different evaluation metrics between the models with and without disease-specific augmentation. From this comparative study, we inferred that with an advantage of handling imbalanced data, the unsupervised approach outperforms the transfer learning model irrespective of whether minIP is used. Furthermore, the increase in sensitivity after minIP of both adversarial and transfer learning models was greater than or equal to 5%, indicating that minIP increases the accuracy of emphysema identification in LDCT data.

V. DISCUSSION AND CONCLUSION

The main goal of this study was to leverage LDCT for automatic detection of emphysema. Using the proposed disease-specific feature learning, unsupervised anomaly detection model, we achieved 12% increase in the area under

Fig. 5. Performance of the proposed network is showed in (a), the receiver operating curve of the network with and without minIP and (b) illustrates the stacked histogram plot of score distribution with and without minIP of the UAD model, where the color red represents abnormal scans and the color blue represents normal scans. We observed that with the application of minIP the score spread has moved towards the lowest and the highest scores indicating improved class separation between the normals and abnormal. For interpretation of the references to colour in the main text, the reader is referred to the web version of this article.

Fig. 6. Top row represents the selected actual anomalous input axial slices and the red bounding box are ground truth annotations by a radiologists. The middle row represents the corresponding post-processed residual images from UAD model. The bottom row represents the corresponding grad-cam images. For interpretation of the references to colour in the main text, the reader is referred to the web version of this article.

TABLE I

PERFORMANCE EVALUATION OF UNSUPERVISED APPROACH AND TRANSFER LEARNING USING DIFFERENT EVALUATION METRICS

Model	Settings	Accuracy	AUC	Sensitivity	Specificity	F1-score
Unsupervised anomaly detection	with minIP	0.88	0.91 $\pm$ 0.03	0.87 $\pm$ 0.02	0.90 $\pm$ 0.01	0.86 $\pm$ 0.02
	without minIP	0.80	0.79 $\pm$ 0.02	0.85 $\pm$ 0.03	0.76 $\pm$ 0.02	0.79 $\pm$ 0.03
	with minIP	0.65	0.70 $\pm$ 0.02	0.68 $\pm$ 0.04	0.68 $\pm$ 0.03	0.68 $\pm$ 0.02
	without minIP	0.67	0.67 $\pm$ 0.04	0.63 $\pm$ 0.03	0.66 $\pm$ 0.03	0.64 $\pm$ 0.02

[] – represents 95% confidence interval

TABLE II

CALCULATED OVERLAP PERCENTAGE BETWEEN NORMAL AND ABNORMAL SCANS ON S_{test} ASSUMING SAMPLES ARE NORMALLY DISTRIBUTED AND USING DESCRIPTIVE STATISTICS

Setting	μ_n	σ	μ_a	σ	Overlap (%)
With minIP	0.1424	0.1357	0.6705	0.2012	13
Without minIP	0.1524	0.1103	0.7206	0.2159	26

 μ – Mean σ – standard deviation

n - non-emphysematous

a- anomaly class (emphysematous)

TABLE III

OVERVIEW OF EMPHYSEMA DETECTION MODELS IN THE LITERATURE

Authors	Models	Features	Modality	Accuracy	AUROC
Our approach	AAE	Disease-specific augmentation (with miniIP) + AAE	LDCT	88%	0.91
Ahmed et al. [23]	VoxResnet3D	3D Transfer learning classification	HRCT	78.3%	-
Humphries et al. [18]	CNN-LSTM	Trained on sequence of slices from each participant	HRCT	79.6%	-
Hatt et al. [22]	CNN	Trained on set of eight axial slices montage from each participant	HRCT	78.9%	-
Sorensen et al. [19]	Texture based	Local binary patterns and K nearest neighbor (KNN) classifier	LDCT	-	0.71

curve(AUC) compared to without minIP. Thus accurately detecting emphysema in a LDCT dataset which is less optimum compared to HRCT. We observed that application of minIP enabled the model to learn faster (in fewer epochs) by reducing the complexity of the image (suppressing the vessels and high-intensity phenotypes) and facilitated the DL model to focus on disease-specific features. Apart from the improvement in standard performance metrics, the increase in class-separation percentage by 13% proved the importance of using clinical domain knowledge like minIP for automatic emphysema detection. We found that disease-specific augmentation increased the performance of the 2.5D Resnet model by 2%. However, the overall performance of the UAD model was superior to the transfer learning approach. This could be because the transfer learning approach requires large-scale data.

The proposed UAD model was robust to contamination up to 3% in the training dataset. However, our contaminated UAD model still outperforms the other state of art emphysema detection models in LDCT with an average AUC of 0.85 (Table III). By including varying anomalies we evaluated the truly unsupervised scenario, and to our knowledge this is the only study to evaluate this aspect of a GAN model in medical imaging [44, 45].

Chuquicusma et al. [46] used visual Turing test for the qualitative analysis of their GAN-generated CT nodules to check the image quality of GAN-generated images. In our study, to make sure that the GAN model is learning the normal distribution well, we employed visual Turing test, rather than the quantitative Fréchet inception distance (FID)

method which is not suitable in limited data settings [8]. We demonstrated that, given a limited data scenario for training, our UAD model is capable of generating convincingly realistic input images, implying efficient mapping of the normality.

The visualisation techniques such as stacked histogram plot, PPR images and grad-cam accounted for the model's interpretability and instilled confidence in the radiologists. A study by [40] showed that PPR images from their GAN model can be used to quantify brain lesions. However for emphysema, the limitation of PPR images is that they cannot be directly associated with emphysema localization due to the presence of other low attenuation areas (trachea or bronchus), and due to the uncertainty that exists with emphysema CT features (lack of ground truth). Since emphysema quantification was out of the current study's scope, we intend to compare PPR images with CT density measurements as a part of our future work. We observed (Fig. 6) that grad-cam was found to pick up some redundancy (regions other than emphysema). Liu Li et al. pointed out a similar issue for glaucoma detection and used an additional attention based-CNN model to localize the anomaly[47]. We intend to use such technique in the future to improve emphysema localization.

This study is a proof of concept for emphysema detection in LDCT. A preliminary external validation of the UAD model was performed on a subset of NLST data and was found to perform with clinically acceptable performance (0.77 ± 0.06). We intend to carry-out an extensive study using large-scale lung cancer screening dataset in near future. Since our proposed model can be used for classifying diseases with

density variations in CT representation, we would like to apply the model on other lung diseases involving lung opacities and consolidation.

In lung cancer screening, where emphysema detection is not the primary focus, the proposed approach can help classify emphysema and increase lung cancer screening sensitivity while reducing the burden on radiologists and may aid early detection of emphysema.

REFERENCES

- [1] Denise R. Aberle et al. "Reduced lung-cancer mortality with low-dose computed tomographic screening". In: *New England Journal of Medicine* (2011). ISSN: 15334406. DOI: 10.1056/NEJMoa1102873.
- [2] Nanda Horeweg et al. "Detection of lung cancer through low-dose CT screening (NELSON): a prespecified analysis of screening test performance and interval cancers". In: *The lancet oncology* 15.12 (2014), pp. 1342–1350.
- [3] Marjolein A Heuvelmans et al. "Screening for early lung cancer, chronic obstructive pulmonary disease, and cardiovascular disease (the Big-3) using low-dose chest computed tomography". In: *Journal of thoracic imaging* 34.3 (2019), pp. 160–169.
- [4] Lara J Akinbami, Xiang Liu, and American Lung Association. "Trends in COPD (Chronic Bronchitis and Emphysema): Morbidity and Mortality". In: *American Lung Association Epidemiology and Statistics Unit Research and Health Education Division* (2013). ISSN: 19414927.
- [5] Pablo Sanchez-Salcedo et al. "Improving selection criteria for lung cancer screening: The potential role of emphysema". In: *American Journal of Respiratory and Critical Care Medicine* (2015). ISSN: 15354970. DOI: 10.1164/rccm.201410-1848OC.
- [6] Robert P Young and Raewyn J Hopkins. "Chronic obstructive pulmonary disease (COPD) and lung cancer screening". In: *Translational lung cancer research* 7.3 (2018), p. 347.
- [7] Raghavendra Chalapathy and Sanjay Chawla. "Deep Learning for Anomaly Detection: A Survey". In: (2019), pp. 1–50. arXiv: 1901.03407. URL: <http://arxiv.org/abs/1901.03407>.
- [8] Thomas Schlegl et al. "f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks". In: *Medical Image Analysis* (2019). ISSN: 13618423. DOI: 10.1016/j.media.2019.01.010.
- [9] N Muller and H Coxson. "Chronic obstructive pulmonary disease• 4: imaging the lungs in patients with chronic obstructive pulmonary disease". In: *Thorax* 57.11 (2002), p. 982.
- [10] Diana Davletshina et al. "Unsupervised Anomaly Detection for X-Ray Images". In: *arXiv preprint arXiv:2001.10883* (2020).
- [11] Ian J. Goodfellow et al. "Generative adversarial nets". In: *Advances in Neural Information Processing Systems*. 2014.
- [12] Phillip Isola et al. "Image-to-image translation with conditional adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134.
- [13] Geert Litjens et al. "A survey on deep learning in medical image analysis". In: *Medical Image Analysis* 42.1995 (2017), pp. 60–88. ISSN: 13618423. DOI: 10.1016/j.media.2017.07.005. arXiv: 1702.05747.
- [14] Thomas Schlegl et al. "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2017. ISBN: 9783319590493. DOI: 10.1007/978-3-319-59050-9_12. arXiv: 1703.05921.
- [15] Alexis Lacout et al. "Multidetector row CT scan in hypersensitivity pneumonitis: contribution of minimum intensity projection reformation". In: *Journal of Medical Imaging and Radiation Oncology* 55.3 (2011), pp. 291–295.
- [16] Shiro Satoh et al. "Pulmonary emphysema: histopathologic correlation with minimum intensity projection imaging, high-resolution computed tomography, and pulmonary function test results". In: *Journal of computer assisted tomography* 32.4 (2008), pp. 576–582.
- [17] Agnieszka Choromańska and Katarzyna J. Macura. *Role of computed tomography in quantitative assessment of emphysema*. 2012. DOI: 10.12659/PJR.882578.
- [18] Stephen M. Humphries et al. "Deep learning enables automatic classification of emphysema pattern at CT". In: *Radiology* (2020). ISSN: 15271315. DOI: 10.1148/radiol.2019191022.
- [19] L. Sorensen, S. B. Shaker, and M. de Bruijne. "Quantitative Analysis of Pulmonary Emphysema Using Local Binary Patterns". In: *IEEE Transactions on Medical Imaging* 29.2 (2010), pp. 559–569.
- [20] Jie Yang et al. "Texton and sparse representation based texture classification of lung parenchyma in CT images". In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. 2016. ISBN: 9781457702204. DOI: 10.1109/EMBC.2016.7590939.
- [21] Esra Mahsereci Karabulut and Turgay Ibrici. "Emphysema discrimination from raw HRCT images by convolutional neural networks". In: *ELECO 2015 - 9th International Conference on Electrical and Electronics Engineering*. 2016. ISBN: 9786050107371. DOI: 10.1109/ELECO.2015.7394441.
- [22] Charles Hatt et al. "Convolutional neural network based COPD and emphysema classifications are predictive of lung cancer diagnosis". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2018. ISBN: 9783030009458. DOI: 10.1007/978-3-030-00946-5_30.
- [23] Jalil Ahmed et al. "COPD classification in CT images using a 3D convolutional neural network". In: *Infor-*

- matik aktuell. 2020. ISBN: 9783658292669. DOI: 10.1007/978-3-658-29267-6_8. arXiv: 2001.01100.
- [24] Wassim W Labaki et al. "Quantitative emphysema on low-dose CT Imaging of the chest and risk of lung cancer and airflow obstruction: an analysis of the National Lung Screening Trial". In: *Chest* 159.5 (2021), pp. 1812–1820.
- [25] Andreea Udrea and George Daniel Mitra. "Generative Adversarial Neural Networks for Pigmented and Non-Pigmented Skin Lesions Detection in Clinical Images". In: *Proceedings - 2017 21st International Conference on Control Systems and Computer, CSCS 2017*. 2017. ISBN: 9781538618394. DOI: 10.1109/CSCS.2017.56.
- [26] Xiaoran Chen and Ender Konukoglu. "Unsupervised detection of lesions in brain mri using constrained adversarial auto-encoders". In: *arXiv preprint arXiv:1806.04972* (2018).
- [27] Lukas Ruff et al. "Deep Semi-Supervised Anomaly Detection". In: (2019). arXiv: 1906.02694. URL: <http://arxiv.org/abs/1906.02694>.
- [28] Christoph Baur et al. "Autoencoders for Unsupervised Anomaly Segmentation in Brain MR Images: A Comparative Study". In: *arXiv preprint arXiv:2004.03271* (2020).
- [29] Tim Salimans et al. "Improved techniques for training gans". In: *Advances in neural information processing systems*. 2016, pp. 2234–2242.
- [30] Ian J Goodfellow et al. "Generative Modeling Generative Modeling". In: *Annual Plant Reviews* (2016). ISSN: 10495258. DOI: 10.1002/9781118472507.fmatter. arXiv: arXiv:1701.00160v4.
- [31] Martin Arjovsky, Soumith Chintala, and Léon Bottou. "Wasserstein generative adversarial networks". In: *34th International Conference on Machine Learning, ICML 2017*. 2017. ISBN: 9781510855144.
- [32] Ishaan Gulrajani et al. "Improved training of wasserstein gans". In: *Advances in neural information processing systems*. 2017, pp. 5767–5777.
- [33] Alec Radford, Luke Metz, and Soumith Chintala. "Unsupervised Representation learning with Deep Convolutional GANs". In: *International Conference on Learning Representations* (2016). ISSN: 0004-6361. DOI: 10.1051/0004-6361/201527329. arXiv: arXiv:1511.06434v2.
- [34] Samet Akcay, Amir Atapour-Abarghouei, and Toby P. Breckon. "Skip-GANomaly: Skip Connected and Adversarially Trained Encoder-Decoder Anomaly Detection". In: *Proceedings of the International Joint Conference on Neural Networks 2019-July* (2019). DOI: 10.1109/IJCNN.2019.8851808. arXiv: 1901.08954.
- [35] Antonia Creswell et al. "Generative adversarial networks: An overview". In: *IEEE Signal Processing Magazine* 35.1 (2018), pp. 53–65.
- [36] Congying Xia et al. "Early imaging biomarkers of lung cancer, COPD and coronary artery disease in the general population: rationale and design of the ImaLife (Imaging in Lifelines) Study". In: *European Journal of Epidemiology* (2020). ISSN: 15737284. DOI: 10.1007/s10654-019-00519-0.
- [37] David A. Lynch et al. "CT-definable subtypes of chronic obstructive pulmonary disease: A statement of the fleischner society". In: *Radiology* (2015). ISSN: 15271315. DOI: 10.1148/radiol.2015141579.
- [38] Shiro Satoh, Shinichi Ohdama, and Hitoshi Shibuya. "Sliding thin slab, minimum intensity projection imaging for objective analysis of emphysema". In: *Radiation Medicine - Medical Imaging and Radiation Oncology* (2006). ISSN: 02882043. DOI: 10.1007/s11604-006-0045-y.
- [39] Houssam Zenati et al. "Efficient gan-based anomaly detection". In: *arXiv preprint arXiv:1802.06222* (2018).
- [40] Christoph Baur et al. "Deep autoencoding models for unsupervised anomaly segmentation in brain MR images". In: *International MICCAI Brainlesion Workshop*. Springer. 2018, pp. 161–169.
- [41] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017. ISBN: 9781538610329. DOI: 10.1109/ICCV.2017.74. arXiv: 1610.02391.
- [42] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [43] Tero Karras et al. "Training generative adversarial networks with limited data". In: *arXiv preprint arXiv:2006.06676* (2020).
- [44] Laura Beggel, Michael Pfeiffer, and Bernd Bischl. "Robust anomaly detection in images using adversarial autoencoders". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2019, pp. 206–222.
- [45] Amanda Berg, Jörgen Ahlberg, and Michael Felsberg. "Unsupervised Learning of Anomaly Detection from Contaminated Image Data using Simultaneous Encoder Training". In: *arXiv preprint arXiv:1905.11034* (2019).
- [46] Maria J.M. Chuquicusma et al. "How to fool radiologists with generative adversarial networks? A visual turing test for lung cancer diagnosis". In: *Proceedings - International Symposium on Biomedical Imaging*. 2018. ISBN: 9781538636367. DOI: 10.1109/ISBI.2018.8363564.
- [47] Liu Li et al. "Attention based glaucoma detection: A large-scale database and CNN Model". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 10571–10580.