Semi-supervised labelling of chest x-ray images using unsupervised clustering for ground-truth generation

Agughasi Victor Ikechukwu [a,1,*], Murali Srinivasiah [a,2]

[a] Maharaja Institute of Technology Mysore, Belawadi, Naguvanahalli Post, Srirangapatna Tq, 571477, Karnataka, India
[1] victor.agughasi@gmail.com; [2] murali@mitmysore.in
* corresponding author

ARTICLE INFO

Article history
Received August 30, 2023
Revised September 07, 2023
Accepted September 10, 2023
Available online September 12, 2023

Keywords
Chest X-Rays
Semi-supervised classifier
Unsupervised clustering
Ground-truth generation
VinDR-CXR dataset

ABSTRACT

Supervised classifiers require a lot of data with accurate labels to learn to recognize chest X-ray images (CXR). However, manually labeling an extensive collection of CXR images is time-consuming and costly. To address this issue, a method for the semi-supervised labelling of extensive collections of CXR images is proposed leveraging unsupervised clustering with minimum expert knowledge to generate ground truth images. The proposed methodology entails: using unsupervised clustering techniques such as K-Means and Self-Organizing Maps. Second, the images are fed to five different feature vectors to utilize the potential differences between features to their full advantage. Third, each data point gets the label of the cluster's center to which it belongs. Finally, a majority vote is used to decide the ground truth image. The number of clusters created by the method chosen strictly limits the amount of human involvement. To evaluate the effectiveness of the proposed method, experiments were conducted on two publicly available CXR datasets, namely VinDR-CXR and Montgomery datasets. The experiments showed that, for a KNN classifier, manually labeling only 1% (VinDr-CXR), or 10% (Montgomery) of the training data, gives a similar performance as labeling the whole dataset. The proposed methodology efficiently generates ground-truth images from publicly available CXR datasets. To our knowledge, this is the first study to use the VinDr-CXR and Montgomery datasets for ground truth image generation. Extensive experimental analysis using machine learning and statistical techniques shows that the proposed methodology efficiently generates ground truth images from CXR datasets.

1. Introduction

Deep learning has recently been used in conjunction with medical imaging to detect and segment abnormalities [1]. Stanford University researchers proposed ChexNet [2], which demonstrated improved accuracy in diagnosing 14 chest X-ray (CXR) detectable diseases compared to medical professionals. Rapid development in AI has altered every facet of human activity. An example of this product type is DeepMind [3], a pioneer in artificial intelligence (AI) that gained notoriety for its groundbreaking work on AlphaGo and its neural network that learned to play games at a level above human ability. The most skilled human player in the ancient Chinese game (Go) was no match for this computer software.

A high mortality rate and high medical costs are associated with Chronic Obstructive Pulmonary Disorder (COPD), a lung ailment. According to the World Health Organization (WHO) [4], COPD will be the third leading cause of death worldwide by 2030 [5]. Patients in the early stages of COPD are commonly missed because they show no symptoms or mild ones [6]. Most patients are already in

the moderate-to-severe stage when diagnosed, which has far-reaching consequences for their everyday lives and treatment options. As a result, detecting COPD early reduces the likelihood of developing respiratory disorders and reduces associated costs. The value of a prompt diagnosis of chronic obstructive pulmonary disease (COPD) is becoming more widely acknowledged. For COPD diagnosis, spirometry [7] is a must-have tool. However, its usefulness is limited to the first stages of COPD, and it is widely regarded as ineffectual.

Consolidation, atelectasis, lung nodules, and masses are some of the radiological and histological characteristics characterizing [8] , as shown in Fig. 1.

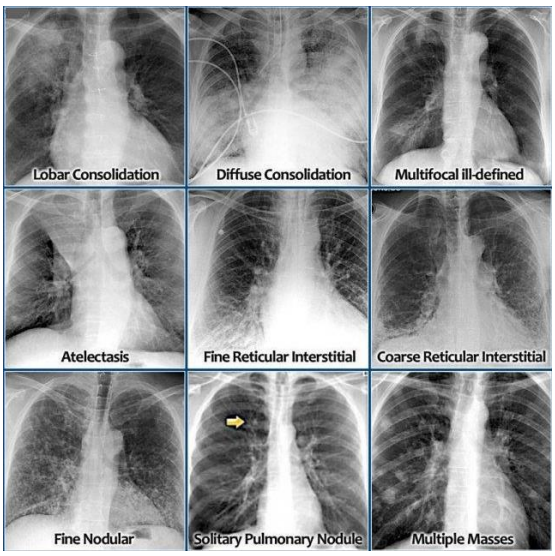

Fig. 1. An illustration of COPD with variability in lesions and infected areas

CXR imaging is the most often used screening method because it allows doctors to quickly and accurately diagnose and treat a wide range of thoracic illnesses, such as emphysema, pleural effusion, pneumonia, and pneumothorax. However, assessing hundreds of radiological samples in real time remains challenging due to the considerable reliance on skilled radiologists' handwritten annotations. In addition, the radiologist has a substantial diagnostic challenge because multiple abnormalities may be visible on a single CXR scan.

Now a spirometry test can validate a preliminary diagnosis of COPD based on CXR pictures. Conventional machine learning techniques were previously used to identify CXR images based on COPD. Feragen et al. [9] analyzed the airway shape of 1996 individuals, 893 of whom had COPD, using a Support Vector Machine (SVM). They attained a maximum accuracy of 65% when classifying patients with COPD. Bodduluri et al. [10] evaluated the performance of the K-Nearest Neighbor (KNN) learning algorithm for diagnosing COPD patients. Texture feature sets with an AUC of 0.89 were the best in the literature [10]. Using Multiple Instance Learning (MIL) techniques, Cheplygina et al. [11] were able to classify cases of chronic obstructive pulmonary disease (COPD) with an AUC of 74.2%.

Classification approaches for large datasets typically use unsupervised learning techniques like K-Means and Self-Organizing Maps (SOMs) [12]; therefore, automatic and high-accuracy labelling processes are crucial. Since this method does not need domain-specific data or metrics, it can be more broadly applied if more of the label discovery process can be automated. While calculating the cost of matching an expression pair, the expression's local and global attributes are considered. In order to classify handwritten digits online, Li et al. [13] propose a codebook mapping that employs agglomerative clustering to group strokes together, along with a mapping function in which the distance between groups of strokes is employed to produce representative labels. Afterward, the results of the learners' efforts are linearly combined and tabulated to implement majority voting rules.

Feature engineering has a fundamental fault in requiring a high level of expert knowledge of the features to be effective. In contrast, thanks to AI advancements, cutting-edge deep-learning approaches have been applied to COPD identification. These techniques could enhance image data

classification without first identifying relevant radiographic features. One of the most powerful deep learning architectures for computer vision is the convolutional neural network (CNN) [14].

To segment the lungs, Ahmad et al. [15] use oriented Gaussian derivatives, thresholding, and Fuzzy C-Means (FCM) clustering. For the JSRT dataset used by the Japanese Society for Radiological Technology, their method achieved an accuracy of above 90% except for the overlap (87%). Considering the scale-dependency of shape and appearance data, deformable model-based approaches to lung segmentation use a joint shape and appearance sparse learning-based framework [16]. Several authors have developed hybrid systems that combine active shape models [17] with other automatic approaches [18] to tackle the complex problem of lung segmentation.

However, because of the wide variety of lung field shapes, the lung borders produced by these traditional segmentation approaches may not be adequate. In addition, these approaches are also unsuccessful when dealing with pulmonary diseases that alter lung texture.

González et al. [19] used their deep CNN models trained on over a thousand COPDGene participants and found that they had an accuracy of 75%. Hatt et al. [20] used a single CNN model to predict lung cancer in the National Lung Screening Trial (NLST) cohort with an accuracy of 77%. Using a cohort study and replicated results from selecting slices in ECLIPSE, Tang et al. [21] obtained an AUC of 0.88. Using a histogram of oriented gradient (HOG) and a pre-trained convolutional neural network (VGGNet) using CXR images was proposed as a hybrid method by Ragab et al. [22]. The noise was eliminated using modified anisotropic diffusion filtering (MADF), and its 99.49% accuracy demonstrated that the proposed method was superior to others.

Score-CAM (Class Activation Mapping) visualization was used by Tahir et al. [23] to discriminate ROI from ordinary CXR images. Results showed a sensitivity of 96.94% from their system, which is encouraging for the future of AI in general. In presenting a CNN-based model for pneumonia classification in CXR images, the work of [24]–[27], showed that it is possible, albeit computationally expensive, to train a deep neural network from scratch on a low-end Computer. They investigated the effects of hyperparameter tuning by trial and error with varying dropout values, achieving greater precision than was previously possible with the most basic methods.

Literary Findings: Researchers have examined the JSRT, ChestX-ray14, CheXpert, and COVID-CXR datasets based on reports in the available literature. K-Means and Fuzzy C-Means clustering and support vector machines (SVM) significantly outperform the state-of-the-art model in accuracy. Nevertheless, when using the recently published clinically-annotated VinDR-CXR dataset, results for COPD ground-truth creation from CXR images are less impressive

To address these issues, we offer a semi-automatic approach using unsupervised clustering with restricted manual annotations that can boost the precision on which ground truths are extracted by taking account of the differences in lung structure. As a result, the proposed model employs fine-tuning by combining various image enhancement methods, such as the Contrast Limited Adaptive Histogram Equalization (CLAHE), to achieve robust results. Fig. 4 depictour proposed methodology.

The primary objectives of the present study can be categorized into three central areas of focus. First, the study seeks to establish a voting mechanism for the governance of accepted labels. Secondly, there is an emphasis on evaluating the efficacy and control of the proposed feature sets. Lastly, the research aims to discern and select an optimal method for data clustering.

Regarding the contributions of this research, several noteworthy advancements and findings are presented. A pioneering method is introduced for generating ground-truth images of lung regions on the VinDr-CXR dataset. Moreover, the incorporation of CLAHE (Contrast Limited Adaptive Histogram Equalization) during the pre-processing phases has been shown to enhance image quality, leading to improved segmentation accuracy for the VinDr-CXR dataset. Additionally, our findings suggest the feasibility of deploying an ensemble model, similar to classification frameworks, for the segmentation of CXR images. A comprehensive comparison, encompassing a wider range of performance metrics than those discussed in the contemporary state-of-the-art models, is provided. Notably, through the adoption of these innovative techniques, our study stands as the first to illustrate results for generating ground truth masks on the VinDr-CXR dataset.

This study is in five subsections. Section one highlights the background and rationale of the study. Section 2 presents related literature on ground truth generation using CXR images. We introduce the

core methodology in Section 3. Section 4 discussed the results of the experiments and how they stacked up against most state-of-the-art models. Finally, section 5 concludes the work, stating the limitations and future directions.

2. Method

The methodology employed in this study can be divided into four core stages, as illustrated in Fig. 5. These stages range from preprocessing the input images to evaluating the classification results.

2.1. Dataset Description

This study uses the clinically validated CXR dataset VinDR-CXR [28]. This database includes over 100,000 chest X-rays from two of Vietnam's largest medical hospitals. It has about 18,000 images as shown in Fig. 2 (15,000 of which serve as training data and 3,000 as test data).

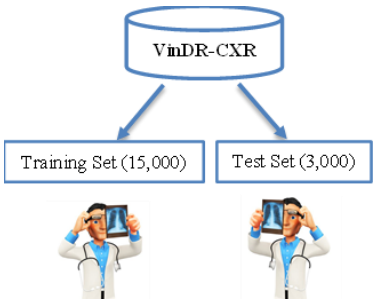

Fig. 2. The dataset distribution withradiological annotations

A team of 17 experienced radiologists carefully labeled each one with 22 local labels of rectangles that enclose abnormalities and six global labels of suspected diseases, as illustrated in Fig. 3. During the training phase, three radiologists labeled each scan independently, whereas a panel of five agreed to identify each scan in the test phase. To ensure conformity to medical standards, labels for the training and validation sets and all de-identified images are freely available in Digital Imaging and Communications in Medicine (DICOM) [29] format.

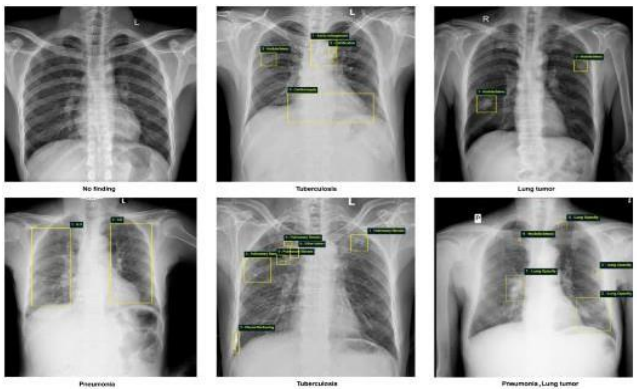

Fig. 3. Samples of the dataset with abnormalitiesenclosed in yellow-rectangular bounding boxes

Fig. 3 shows that no segmented ground truth existed for the CXR images. To this end, we employed a semi-supervised method to build a ground truth after exploring a different dataset, the Montgomery dataset [30], which included validated ground truth. 138 Postero-Anterior (PA) CXR in the archive; 80 are considered to be within normal limits, while the remaining 58 exhibit abnormalities suggestive of tuberculosis. These CXR images, known as the Montgomery dataset, were collected as part of a campaign to combat tuberculosis in Montgomery, Alabama, USA. All images are provided in DICOM format after being thoroughly purged of any personally identifying information.

2.2. Pre-processing of Images

In contrast to the consistency of the Montgomery CXR dataset (all images were 4982x4020), the VinDr-CXR dataset had a wide variation in image size, with most images being within the range of 2500x2500. To use the abundance of DICOM images, we coded a Python script to transform DICOM files into their matching PNG files automatically. To facilitate quicker training, we reduced the

images' resolution to 224 × 224, which proved effective when segmenting microscopic white blood cells (WBC) [31] for clinical diagnosis following normalization. It ensures the intensity of each pixel is between 0 and 1; 0 is completely black and 1 is entirely white. Using a numeric scale from 0 to 1, we may identify a variety of shades of grey. When merging datasets from several sources, it is essential to normalize the pixel values to a standard scale.

Further, we used the Contrast-Limited Adaptive Histogram Equalization (CLAHE) to enhance the lung areas, and the results are in Fig. 4.

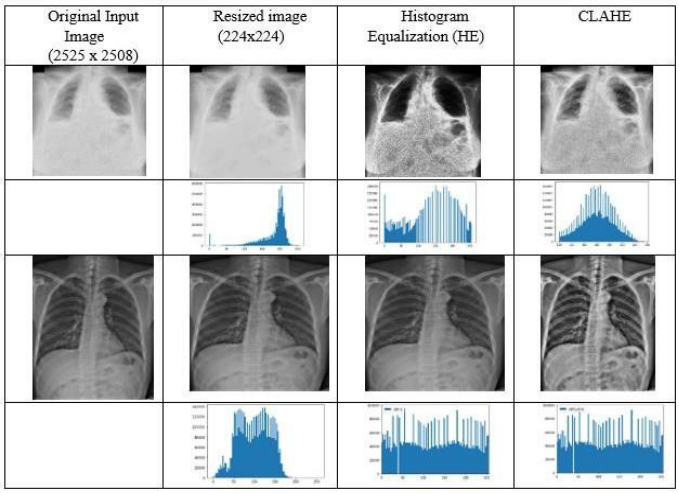

Fig. 4. The result of the processing operations on the VinDR-CXR images

While Histogram Equalization (HE) can be helpful, it resulted in an overamplification of the intensity levels in our experiment. Because of this, CLAHE was used as it reduced the noise amplification. On the other hand, local contrast amplification is controlled by the gradient of the transformation function in CLAHE.

2.3. Proposed Methodology

The proposed method consists of 4 intermediate steps from the input image to evaluation as illustrated in Fig. 5.

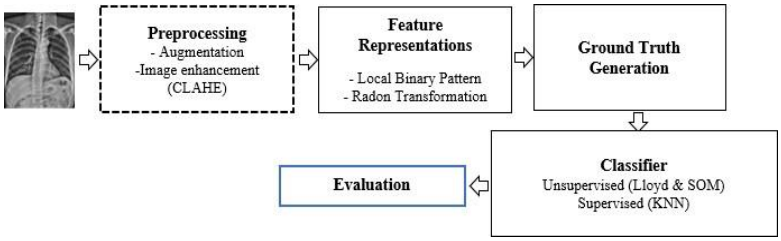

Fig. 5. The result of the processing operations on the VinDR-CXR images

First, to maintain uniformity, the images were pre-processed. Then, applied data augmentation with variations of scaling and rotations followed by the CLAHE for image enhancement. The parameters chosen are in Table 1.

Table 1. Parameters for Image Augmentation on the VinDR-CXR Dataset

Method	Default	Augmented
HorizontalFlip	None	True (p = 0. 5)
VerticalFlip	None	True (p = 0. 5)
Rescale (Normalization)	-	1./255
Zoom range	-	0.25
Rotation (°)	-	60, 90 & 120
x-Shift, y-Shift	None	[- 0.1, + 0. 1]
x-Scale, y-Scale	None	[0.75, 1.25]
Adjusted image	1024 x 1024	224 4

2.3.1. Feature Representations

The various feature spaces must work together to take advantage of the method's potential fully. However, such a complement cannot be determined beforehand. As a result, we arbitrarily chose traits from a set often employed in the literature [32]. Some are thought to be quite effective, while others are not. Here are a few: A local binary pattern (LBP), the Radon transform, and an encoder network. We represent the Grayscale image of dimensions as follows.

$$Ia * Ib \tag{1}$$

Where Ia, and Ib represents the dimensions in the x and y coordinates

2.3.1.1. Local Binary Patterns (LBP)

In order to recognize malignant breast tumors, the local texture was used in conjunction with local binary patterns (LBP) [33], which proved effective. The LBP's observation of the local neighbourhood yields a good result despite the heterogeneity of mammographs in terms of discriminatory power. LBP is a neighborhood-based local description of an image. For example, using the LBP approach, an extreme edge detector can find all edges in an image.

2.3.1.2. Radon Transform(RT)

The Radon transform [34] calculates an image's projections in specified directions. Multiple beams traveling in parallel can have their line integrals calculated with the help of the Radon function. The distance between each pair of beams is one pixel. The Radon function uses several projections of the image from various angles about the image's center to represent the image. The projections are further processed to generate feature vectors for classification purposes. The feature vectors generated by these methods are then fed into the classifiers for training and evaluation.

2.3.2. Ground-truth Generation

2.3.2.1. Encoder-Decoder Network

Hinton and Salakhutdinov [35] first proposed encoder networks, a subset of deep learning architectures. Such an encoder network is data-driven, unlike the traditional approaches such as Principal Component Analysis (PCA) [36]. An illustration of the encoder-decoder model for ground-truth generation is as in Fig. 6

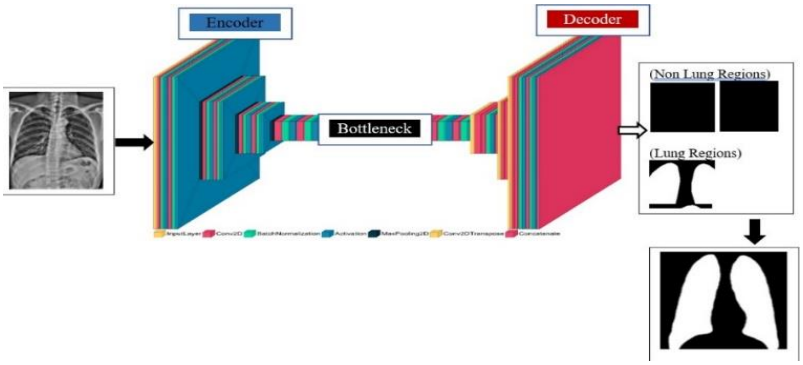

Fig. 6. The Encoder-Decoder block diagram for feature dimensionality reduction

2.3.3. Classifiers

Given a classification problem with $Nclass > 2$ classes, the notation for each category is Cc, $i \xi \{1, 2, 3, \ldots Nc\}$. Let us say we have a classifier Cf that acts as a set of Np patterns, each of which is an input pattern with Nclass characteristics. The patterns are placed in the class denoted by True Positive (TP), following classification, denoted by True Negative (TN). Then, the classifier's accuracy is defined by equation 2 :

$$Accuracy = \frac{Tp+Tn}{TP+TN+FP+FN} \tag{2}$$

Thus, supervised learning is used to acquire the latest spatial projection. The outline for such an encoder system is shown in Fig. 5. The outputs of the bottleneck layer are used as the features under

consideration. The thickness of each layer determines the features dimensionality. Our prior work [25] demonstrated the usefulness of such capabilities.

Specifically, we define combination methods like majority voting and consensus voting for an ensemble of classifiers Ec, $k\ \xi\ \{1, 2, 3, \ldots K\}$ where K is the total number of classifiers. A consensus is reached when at least k experts agree that an input pattern x belongs to a particular category. By majority vote, we designate the answer where k = K, and by consensus, we use the following notation:

$$f(k) = \{\frac{\frac{k}{2}+,if\ K\geq 0(even)}{(K+1)2,if\ K<0(odd)} \quad (3)$$

For performance evaluation, we will refer to Nexp, the number of patterns an expert (Exp) combination assigned a class, and Nnot, for which no class assignment was made $Nexp + Nnot = Np$. We label the number of correctly classified Nexp patterns as Nc, and the number of wrongly categorized ones as Nr. We use the following metrics to evaluate the combined classifiers' performance:

$$Exp = \frac{N_{exp}}{N_p} \quad (4)$$

$$Acc(Correct) = \frac{N_c}{N_p} \quad (5)$$

$$Acc = \frac{N_c}{N_{exp}} \quad (6)$$

Where "Exp" represents the number of judgments made by the ensemble of classifiers. Accuracy, in this context, is denoted by Exp, whereas the proportion of correct detections among all $Nexp$ judgments is denoted by $Acc_{correct}$.

2.3.3.1. Unsupervised Clustering Approaches

The differences and similarities between two cutting-edge unsupervised clustering methods were examined. First, the generic Voronoi technique (a loosely used term for K-Means) is a way to find k-partitions among data points that are both well-shaped and uniformly distributed [37].

After each iteration, the points are re-divided based on their distance from the estimated centroid. Second, we employ an unsupervised neural network called a Self-Organizing Map (SOM) to create a two- dimensional input data representation. Using an unsupervised learning neural network trained with a competitive learning algorithm, SOMs, a technique invented by Professor Teuvo Kohonen [38] in the early 1980s, generate subspaces. There is a rebalancing of neuron weights based on their proximity to cells that have been declared winners (i.e., neurons that most closely resemble a sample input). Using multiple input data sets during training, clusters of comparable neurons are formed, and clusters of different neurons are eliminated.

2.3.3.2. Supervised Classification

One of the first supervised classification approaches, the K-Nearest Neighbour (KNN) classifier, was proposed by Fix and Hoges [39]. Most of the KNN of an unknown data sample is used to determine its class label when using an annotated dataset. The KNN classifier is a specific example of the classifier above for k = 1. The approach excels at classification problems and has some desirable properties, including being easy to use, efficient, non-parametric, and fast [40]. However, substantial difficulties arise from adjusting its parameters, such as the neighbourhood size (k) used in the topological representation quality, which can significantly impact the outcome [41].

3. Results and Discussion

This section presents the experimental results and discusses their implications. The experiments were designed to evaluate the effectiveness of the proposed methodology in generating accurate ground-truth labels and in the classification of CXR images.

3.1. Evaluation Method

Several metrics based on quality and quantity have been proposed to evaluate clustering methods objectively. These metrics can be used to compare different clustering algorithms' performance and determine the optimal number of clusters. One common metric used to evaluate the performance of clustering algorithms is cluster compactness. Cluster compactness measures how closely related the data points within a cluster are to each other. In other words, it measures how tightly the data points are clustered together. The standard definition of vector variance, as shown in equation 7, is one of the measures used to evaluate cluster compactness.

$$Var(X) = \sqrt{\frac{1}{M}\sum_{i}^{M} = 1d^2(x,\mu)} \tag{7}$$

Where $d(x,y)$ = the distance between two vectors x and y in X.

$\mu = \frac{1}{M}\sum_i xi$ = mean of the clustered vectors. Thus, using equation 7, we define the cluster compactness as in equation 8:

$$\text{Compactness} = \frac{1}{Kx}\sum_{1}^{kc}\frac{Var(Ci)}{Var(X)} \tag{8}$$

Where Var is the Variance of "C_i" and "x" respectively

When evaluating clustering algorithms, it's essential to use compactness measures that consider both interclass and intraclass variations. This is because interclass variations measure the differences between different clusters, while intraclass variations measure the differences within a cluster. Therefore, it is important to integrate both variations to get a complete picture of the clustering algorithm's performance. While interclass and intraclass variations are essential, choosing which to prioritize does not significantly affect the evaluation outcomes.

However, it is essential to use a smaller number of clusters when using a labelling approach, which involves assigning labels to the clusters. This is because using fewer clusters can increase the reliability of the labelling approach by making it easier to assign meaningful labels to the clusters. When using a large number of clusters, it can be challenging to assign meaningful labels, and the labels may not accurately reflect the underlying patterns in the data.

3.2. The Labelling Strategy

This section provides a pseudocode presentation of the method for the labelling technique as show in Fig. 7. The method provides a deterministic way to label the patterns with minimal human involvement, allowing for the automatic clustering of M-representations of $Input(xinp)$ via unsupervised clustering strategies with K-clusters. This is done by examining patterns $\boldsymbol{xinp}$, $1 \leq inp \leq Np$ represented by N features in $\boldsymbol{Fn}$, $1 \leq n \leq N$.

Algorithm 1: Strategy for computing the majority vote

```
In t al ze po nts:                    ∀i ∈ [1,n]
   Let Feature: Fn, ∀j ∈ [1,K] be any ||Fn|| ≪ r
m = the number of clusters n the method Cm
For n = 1,k do:
Cluster Cm(Fn) be cluster centers Ct,T ∈ [1,l]
Label cluster centers & expand to all clustered ne ghbours
For m = 1,l do:
For o = 1, K do:
Label (fl j) ← Label (fl )     n Fn vector space endFor
If Label (fl j) = K, then
Label (fl ) ← max(Label (fl j))
else
Label (fl ) ← Null
```

Fig. 7. Presentation of the method for the labelling technique

The points in each Fn representation are grouped into k-clusters, with the centroid of each cluster denoted by the symbol Ct. The centroids stand in for features, and are not shown to a human expert in order to have them assign a label on the spot. Accordingly, proximal point is chosen for each Ct, and the human expert manually annotates each original data point. Each point then takes on the label of its parent cluster in the feature representation.

Voting determines the value labelled for a given data point. The deciding factor is the total number of favourable votes cast for all possible representations of the Fj features [41], [42]. Consensus and majority voting methods are also taken into account.

While the time spent on clustering may be ample, the time spent on ground-truth data generation is negligible in comparison to the total amount of clusters that needs labelling. More exact labels need to be sent back to the data points as more cluster centers are to be identified. An appropriate trade-off between the number of feature sets (M) and the number of clusters (K) should be maintained to reduce the need for human knowledge without sacrificing annotation accuracy. When designing our experiment, we examined the k-means and SOM algorithms so that we could precisely regulate the number of clusters.

3.3. Experimental Results

The proposed methodology was evaluated on the VinDR-CXR dataset, split using the 80:20 rule corresponding to training and testing, respectively, and trained for 50 epochs. For the validation set, 20% of the VinDR-CXR dataset was utilised for hyperparameter tuning and improving the model's performance.

The model was trained on a Windows 11 PC with an Intel(R) Pentium(R) Core i7 8th Generation CPU clocked at 2.30GHz and a 6GB GeForce GTX 1060 graphics accelerator card, and the results demonstrate high accuracy in ground-truth label generation and CXR image classification. The classifiers maintained an accuracy of over 98%, which is comparable to manual labeling by experienced radiologists.

3.3.1. Results of the Clustering Performance

The results of the proposed ensemble method is benchmarked on the VinDr-CXR and Montgomery datasets.

Table 2 displays the input features' accuracy and size when considering label information. For a KNN classifier (k = 1), those calculations show the optimal distance between labels and predictions using the Euclidean distance formula. The proposed approach aims to demonstrate that equivalent results can be generated with significantly less involvement of human annotators.

Table 2. KNN (K=1) Classification Accuracy for VinDR-CXR (17 Labels) and Montgomery Datasets(5 Labels)

KNN Feature type	VinDR-CXR		Montgomery	
	Accuracy (%)	*Num. of Features*	*Accuracy (%)*	*Num. of Features*
Encoders	96.20	256	96.80	256
LBP	95.14	224	95.14	224
Pixels	97.10	50, 176 (224 x 224)	98.20	50, 176 (224 x 224)
Radon	97.20	700	98.20	720

In this case of VinDr-CXR, with an accuracy of 97.2%, the primary feature exceeds the other features. Interestingly, the Montgomery dataset obtained an accuracy of 98.2% using the Radon transform. The local binary pattern is marginally less effective (95.14%) than those obtained by the networks for VinDR-CXR. The LBP provides averages for both sets of data. Raw images and Radon transform results for VinDR-CXR and Montgomery are used as a starting point, despite the nearest neighbor performing slightly better for bigger k-values. The ground truth generation results using both datasets are in Fig.8.

Dataset	Input Image(224 x 224)	After CLAHE	GroundTruth (Provided)	GroundTruth (Generated)
VinDr-CXR			Nil	
			Nil	
Montgomery				
				

Fig. 8. Results of Ground-Truth Generation on both VinDR-CXR and Montgomery datasets

The Montgomery dataset has ground-truth images provided by seasoned radiologists, unlike the VinDR-CXR dataset. Our approach provided near-human annotations with semi-supervised clustering for comparative analysis. As in the case of VinDR-CXR, only bounding-boxes were available. Thus, benchmarking was done in consultations with expert radiologists, and the results are presented in Fig. 6.

Table 3 lists the compactness measurements (see Eq. (8)) obtained by several approaches. Comparing VinDR-CXR and Montgomery's raw pixel data (with 100, 200, and 300 cluster centers in mind) reveals a consistent pattern. The greater the number of possible clusters, the tighter the clusters will be. Similar tendencies are visible for a variety of other characteristics.

Table 3. Compactness Results of K-Means and SOM Using only raw Pixel Values

Adopted methods	**VinDR-CXR**			**Montgomery**		
	# of Clusters			***# of Clusters***		
	100	***200***	***300***	***100***	***200***	***300***
K-means	91.67	88.56	85.41	96.80	95.23	91.30
SOM	93.14	91.32	89.20	95.14	90.12	88.50

The Self-Organizing Map (SOM) is a clustering algorithm that arranges clusters in a fixed spatial arrangement. While the SOM may not be as effective as K-Means in some cases, it performs well when the number of clusters is enormous. This is because the SOM's rigidity in the spatial arrangement of neurons allows it to handle a large number of clusters more effectively than other clustering algorithms.

3.3.2. Results of the Labelling Performance

A decision of complete majority voting was investigated to report findings on the correctness of the labelling. The portion of the original data for which the votes indicate that both labels are accurate was reported. When three out of five methods agree on an image's label (majority voting) or when all five methods agree on an image's label (average), we refer to this proportion as the accuracy (AccCorrect).

Consensus voting is a powerful constraint. As shown in Table 4, the simple majority criteria is met for each feature and clustering method far more often than the absolute majority. Once the voting schemes are described, and the samples chosen for the new training set are known, it was clear that the majority vote yields the highest accuracy. If a majority favors a particular sample's label, that sample and its label can be more confidently accepted as representative of its category. However, for underrepresented classes, some classes might be drawn by larger ones, leading to incorrect labelling and a decreased chance of selection in the voting process.

A KNN (k = 1) classification was performed after the new training sets were constructed using the votes, and the results were compared to those in Table 3 for reference. K-means with a maximum of 17 expertly radiology-labeled samples yield a 99.98% score for a majority vote, comparable to the 98.20% observed for considering only pixel values from the 18,000 PA-CXR images in the VinDR-CXR dataset.

Table 4. Values on K-Means, and SOM Techniques using the VinDR-CXR and Montgomery datasets."*No.*" Represents the Number of Human Annotators

Method	**VinDR-CXR**						**Montgomery**						**No.**
	Majority Voting			***Unanimous votes***			***Majority voting***			***Unanimous votes***			
	Train		***Test***	***Train***		***Test***	***Train***		***Test***	***Train***		***Test***	
	Exp	***Acc(w)***	***Acc***	***Exp***	***Acc(w)***	***Acc***	***Exp***	***Acc(w)***	***Acc***	***Exp***	***Acc(w)***	***Acc***	
K-means	96.24	95.40	99.20	55.52	96.45	90.38	38.52	88.41	98.45	51.32	89.47	88.64	10
	95.30	92.60	**99.98**	44.92	97.15	91.62	47.92	88.92	**99.15**	48.62	93.42	85.32	17
SOM	93.20	95.30	98.30	39.78	95.45	92.38	49.78	89.68	98.49	45.78	95.68	88.41	10
	91.50	93.20	**99.52**	41.34	93.85	93.30	47.34	89.56	**99.98**	47.34	92.56	88.54	17

3.3.3. Results of Classification Performance

VinDR-CXR and Montgomery's accuracy was peaked at 99.98% with 3,000 test images and 99.80% with 50 test images respectively.

Table 5 shows that more than 95% accuracy is achievable with the same setup but with fewer labels. Again, the results require only a little number of human-labeled samples.

The Monte Carlo technique [43] was used to generate select samples from the training set and associated labels, demonstrating its efficacy. In addition, pixel values (VinDR-CXR) and Radon (Montgomery) are applied to KNN to provide a direct comparison. The set sizes evaluated range from 30 to 80.

Table 5. Values on KNN Accuracy with Standard Deviations (SD) using the VinDR-CXR and Montgomery Datasets

No of Samples	**VinDR-CXR**		**Montgomery**	
	Random Samples	**ProposedMethod**	**RandomSamples**	**ProposedMethod**
30	80+/-0.90	**89.02**	79+/-1.90	**89.22**
50	85+/-0.81	**93.51**	75+/-1.81	**84.51**
80	89+/-0.91	**97.60**	79+/-0.98	**88.60**

Table 5 shows that, for VinDR-CXR, the reported scores are up to 7.9% lower than ours when utilizing the same number of labels. The net gain for the Montgomery data set is 8.3%, demonstrating that the proposed strategy is superior to solutions with 224 and 256 features, respectively. The significant benefit entails the small number of images that need to be generated, while the feature variety and clustering algorithms can significantly influence the results.

Compared to VinDR-CXR, Montgomery CXR images offer a speedup of roughly 90 times, while VinDR-CXR's is just 40 times if we assume it takes 30 seconds to label one CXR image. Having fewer images to label can also reduce the money spent on having an expert radiologist label them

4. Conclusion

This research introduces a semi-supervised automatic ground-truth generation technique that necessitates minimal human intervention. Each image in the dataset was mapped into one of five distinct feature spaces for this purpose. The management of the number of nodes in each cluster was meticulously conducted through K-means clustering and Self-Organizing Maps (SOM) during the clustering of the feature spaces. Subsequently, the peripheral elements of the cluster adopted the label closest to the cluster's center, as validated by a seasoned radiologist. Multiple voting techniques were employed to determine the final labels after the completion of clustering in each region.

Firstly, under the KNN approach, the newly discovered labels were compared with pre-existing labels and randomly selected samples (with the same expected number of labels). The second phase of the results involved comparing the found labels with a comprehensive dataset, yielding closely aligned results. While labeling the entire dataset might entail significant time and financial investment, the proposed technique delivers comparable outcomes with significantly less reliance on human annotators. The approach accelerates the process and reduces costs, maintaining over 98% accuracy.

To further validate the effectiveness of the proposed methodology, it is essential to evaluate it against a sizeable and complex clinically validated dataset, ideally encompassing a considerable number of potential DR cases. Future work plans include assessing new datasets using DenseNet and VGG19, along with additional augmentation techniques.

Declarations

Author contribution. Agughasi Victor Ikechukwu: Conceptualization, Investigation of challenges, Data collection, Design, Writing- original draft. Writing- review and editing, Analysis and Interpretation of results. Murali S: Supervision, Investigation of challenges, and Draft manuscript verification

Funding statement. This research received no external funding.

Conflict of interest. The authors declare no conflict of interest.

Additional information. No additional information is available for this paper.

Data and Software Availability Statements

The data used for this study were obtained from the clinically validated PhysioNet Database (https://physionet.org/content/vindr-cxr/1.0.0/), and https://data.lhncbc.nlm.nih.gov/public/Tuberculosis-Chest-X-Ray-Datasets/Montgomery-County-CXR-Set/MontgomerySet/index.html)

References

[1] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017, doi: 10.1016/j.media.2017.07.005.

[2] P. Rajpurkar *et al.*, "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," *arXiv Comput. Vis. Pattern Recognit.*, pp. 1–7, Nov. 2017. [Online]. Available at: https://arxiv.org/abs/1711.05225.

[3] D. Silver *et al.*, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016, doi: 10.1038/nature16961.

[4] R. Sivaramakrishnan *et al.*, "Comparing deep learning models for population screening using chest radiography," in *Medical Imaging 2018: Computer-Aided Diagnosis*, Feb. 2018, vol. 10575, p. 49, doi: 10.1117/12.2293140.

[5] C. D. Mathers and D. Loncar, "Projections of Global Mortality and Burden of Disease from 2002 to 2030," *PLoS Med.*, vol. 3, no. 11, p. e442, Nov. 2006, doi: 10.1371/journal.pmed.0030442.

[6] N. Zhong *et al.*, "Prevalence of Chronic Obstructive Pulmonary Disease in China," *Am. J. Respir. Crit. Care Med.*, vol. 176, no. 8, pp. 753–760, Oct. 2007, doi: 10.1164/rccm.200612-1749OC.

[7] J. Wanchaitanawong *et al.*, "A Predictive Model using Artificial Intelligence on Chest Radiograph in Addition to History and Physical Examination to Diagnose Chronic Obstructive Pulmonary Disease," *J. Med. Assoc. Thail.*, vol. 104, no. Suppl. 4, pp. 79–87, Oct. 2021, doi: 10.35755/jmedassocthai.2021.S04.00049.

[8] "Copd Chronic Obstructive Lung Disease (Copd) / Emphysema." https://www.stritch.luc.edu/lumen/meded/radio/curriculum/medicine/emphysema.htm.

[9] A. Feragen *et al.*, "Geometric Tree Kernels: Classification of COPD from Airway Tree Geometry," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7917 LNCS, Springer, Berlin, Heidelberg, 2013, pp. 171–183, doi: 10.1007/978-3-642-38868-2_15.

[10] S. Bodduluri, J. D. Newell, E. A. Hoffman, and J. M. Reinhardt, "Registration-Based Lung Mechanical Analysis of Chronic Obstructive Pulmonary Disease (COPD) Using a Supervised Machine Learning Framework," *Acad. Radiol.*, vol. 20, no. 5, pp. 527–536, May 2013, doi: 10.1016/j.acra.2013.01.019.

[11] V. Cheplygina, L. Sorensen, D. M. J. Tax, J. H. Pedersen, M. Loog, and M. de Bruijne, "Classification of COPD with Multiple Instance Learning," in *2014 22nd International Conference on Pattern Recognition*, Aug. 2014, pp. 1508–1513, doi: 10.1109/ICPR.2014.268.

[12] J. Faigl and G. A. Hollinger, "Autonomous Data Collection Using a Self-Organizing Map," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 5, pp. 1703–1715, May 2018, doi: 10.1109/TNNLS.2017.2678482.

[13] J. Li, H. Mouchere, and C. Viard-Gaudin, "Reducing Annotation Workload Using a Codebook Mapping and Its Evaluation in On-Line Handwriting," in *2012 International Conference on Frontiers in Handwriting Recognition*, Sep. 2012, pp. 752–757, doi: 10.1109/ICFHR.2012.259.

[14] K.-C. Yuan, L.-W. Tsai, K. Lai, S.-T. Teng, Y.-S. Lo, and S.-J. Peng, "Using Transfer Learning Method to Develop an Artificial Intelligence Assisted Triaging for Endotracheal Tube Position on Chest X-ray," *Diagnostics*, vol. 11, no. 10, p. 1844, Oct. 2021, doi: 10.3390/diagnostics11101844.

[15] W. S. H. M. Wan Ahmad, W. M. D. W Zaki, and M. F. Ahmad Fauzi, "Lung segmentation on standard and mobile chest radiographs using oriented Gaussian derivatives filter," *Biomed. Eng. Online*, vol. 14, no. 1, p. 20, Dec. 2015, doi: 10.1186/s12938-015-0014-8.

[16] Y. Shao, Y. Gao, Y. Guo, Y. Shi, X. Yang, and D. Shen, "Hierarchical Lung Field Segmentation With Joint Shape and Appearance Sparse Learning," *IEEE Trans. Med. Imaging*, vol. 33, no. 9, pp. 1761–1780, Sep. 2014, doi: 10.1109/TMI.2014.2305691.

[17] D. K. Iakovidis, M. A. Savelonas, and G. Papamichalis, "Robust model-based detection of the lung field boundaries in portable chest radiographs supported by selective thresholding," *Meas. Sci. Technol.*, vol. 20, no. 10, p. 104019, Oct. 2009, doi: 10.1088/0957-0233/20/10/104019.

[18] B. van Ginneken and B. M. ter Haar Romeny, "Automatic segmentation of lung fields in chest radiographs," *Med. Phys.*, vol. 27, no. 10, pp. 2445–2455, Oct. 2000, doi: 10.1118/1.1312192.

[19] G. González *et al.*, "Disease Staging and Prognosis in Smokers Using Deep Learning in Chest Computed Tomography," *Am. J. Respir. Crit. Care Med.*, vol. 197, no. 2, pp. 193–203, Jan. 2018, doi: 10.1164/rccm.201705-0860OC.

[20] C. Hatt, C. Galban, W. Labaki, E. Kazerooni, D. Lynch, and M. Han, "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)*, vol. 11040 LNCS, Springer Verlag, 2018, pp. 302–309, doi: 10.1007/978-3-030-00946-5_30.

[21] L. Y. W. Tang, H. O. Coxson, S. Lam, J. Leipsic, R. C. Tam, and D. D. Sin, "Towards large-scale case-finding: training and validation of residual networks for detection of chronic obstructive pulmonary disease using low-dose CT," *Lancet Digit. Heal.*, vol. 2, no. 5, pp. e259–e267, May 2020, doi: 10.1016/S2589-7500(20)30064-9.

[22] D. A. Ragab, M. Sharkas, S. Marshall, and J. Ren, "Breast cancer detection using deep convolutional neural networks and support vector machines," *PeerJ*, vol. 7, no. 1, p. e6201, Jan. 2019, doi: 10.7717/peerj.6201.

[23] A. M. Tahir *et al.*, "Deep Learning for Reliable Classification of COVID-19, MERS, and SARS from Chest X-ray Images," *Cognit. Comput.*, vol. 14, no. 5, pp. 1752–1772, Sep. 2022, doi: 10.1007/s12559-021-09955-1.

[24] A. Victor Ikechukwu, S. Murali, R. Deepu, and R. C. Shivamurthy, "ResNet-50 vs VGG-19 vs training from scratch: A comparative analysis of the segmentation and classification of Pneumonia from chest X-ray images," *Glob. Transitions Proc.*, vol. 2, no. 2, pp. 375–381, Nov. 2021, doi: 10.1016/j.gltp.2021.08.027.

[25] A. Victor Ikechukwu and M. S, "CX-Net: an efficient ensemble semantic deep neural network for ROI identification from chest-x-ray images for COPD diagnosis," *Mach. Learn. Sci. Technol.*, vol. 4, no. 2, p. 025021, Jun. 2023, doi: 10.1088/2632-2153/acd2a5.

[26] V. I. Agughasi, Y. DK, and S. M. Das, "Early Prognosis of Heart Failure from Clinical Symptoms using K-Means and Naïve Bayes Algorithms - Peer-reviewed Journal," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 9, no. 7, pp. 55–61, 2020. [Online]. Available at: https://ijarcce.com/papers/early-prognosis-of-heart-failure-from-clinical-symptoms-using-k-means-and-naive-bayes-algorithms/.

[27] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, "Explainable Deep Learning for Pulmonary Disease and Coronavirus COVID-19 Detection from X-rays," *Comput. Methods Programs Biomed.*, vol. 196, p. 105608, Nov. 2020, doi: 10.1016/j.cmpb.2020.105608.

[28] H. Q. Nguyen *et al.*, "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations," *Sci. Data*, vol. 9, no. 1, p. 429, Jul. 2022, doi: 10.1038/s41597-022-01498-w.

[29] F. Rahimi and H. Rabbani, "A dual adaptive watermarking scheme in contourlet domain for DICOM images," *Biomed. Eng. Online*, vol. 10, no. 1, p. 53, Jun. 2011, doi: 10.1186/1475-925X-10-53.

[30] S. Candemir *et al.*, "Lung Segmentation in Chest Radiographs Using Anatomical Atlases With Nonrigid Registration," *IEEE Trans. Med. Imaging*, vol. 33, no. 2, pp. 577–590, Feb. 2014, doi: 10.1109/TMI.2013.2290491.

[31] A. V. Ikechukwu and S. Murali, "i-Net: a deep CNN model for white blood cancer segmentation and classification," *Int. J. Adv. Technol. Eng. Explor.*, vol. 9, no. 95, pp. 1448–1464, Oct. 2022, doi: 10.19101/IJATEE.2021.875564.

[32] S. Sun and R. Zhang, "Region of Interest Extraction of Medical Image based on Improved Region Growing Algorithm," in *Proceedings of the 2017 International Conference on Material Science, Energy and Environmental Engineering (MSEEE 2017)*, Aug. 2017, pp. 471–475, doi: 10.2991/mseee-17.2017.87.

[33] M. Wei *et al.*, "A Benign and Malignant Breast Tumor Classification Method via Efficiently Combining Texture and Morphological Features on Ultrasound Images," *Comput. Math. Methods Med.*, vol. 2020, pp. 1–12, Oct. 2020, doi: 10.1155/2020/5894010.

[34] K. E. Barner, "Region of interest identification in collimated x-ray images utilizing nonlinear preprocessing and the Radon transform," *J. Electron. Imaging*, vol. 14, no. 3, p. 033011, Jul. 2005, doi: 10.1117/1.2005042.

[35] G. E. Hinton and R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science (80-.).*, vol. 313, no. 5786, pp. 504–507, Jul. 2006, doi: 10.1126/science.1127647.

[36] C. Chen, H. Seo, C. H. Jun, and Y. Zhao, "Pavement crack detection and classification based on fusion feature of LBP and PCA with SVM," *Int. J. Pavement Eng.*, vol. 23, no. 9, pp. 3274–3283, Jul. 2022, doi: 10.1080/10298436.2021.1888092.

[37] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010, doi: 10.1016/j.patrec.2009.09.011.

[38] T. Kohonen, "The self-organizing map," *Proc. IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990, doi: 10.1109/5.58325.

[39] E. Fix and J. L. Hodges, "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties," *Int. Stat. Rev. / Rev. Int. Stat.*, vol. 57, no. 3, p. 238, Dec. 1989, doi: 10.2307/1403797.

[40] A. Torralba, R. Fergus, and W. T. Freeman, "80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1958–1970, Nov. 2008, doi: 10.1109/TPAMI.2008.128.

[41] L. I. Kuncheva, *Combining Pattern Classifiers*. Wiley, p. 350, Jul. 2004, doi: 10.1002/0471660264.

[42] J. Kittler and F. M. Alkoot, "Sum versus vote fusion in multiple classifier systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 1, pp. 110–115, Jan. 2003, doi: 10.1109/TPAMI.2003.1159950.

[43] D. Sarrut, A. Etxebeste, E. Muñoz, N. Krah, and J. M. Létang, "Artificial Intelligence for Monte Carlo Simulation in Medical Physics," *Front. Phys.*, vol. 9, p. 738112, Oct. 2021, doi: 10.3389/fphy.2021.738112.