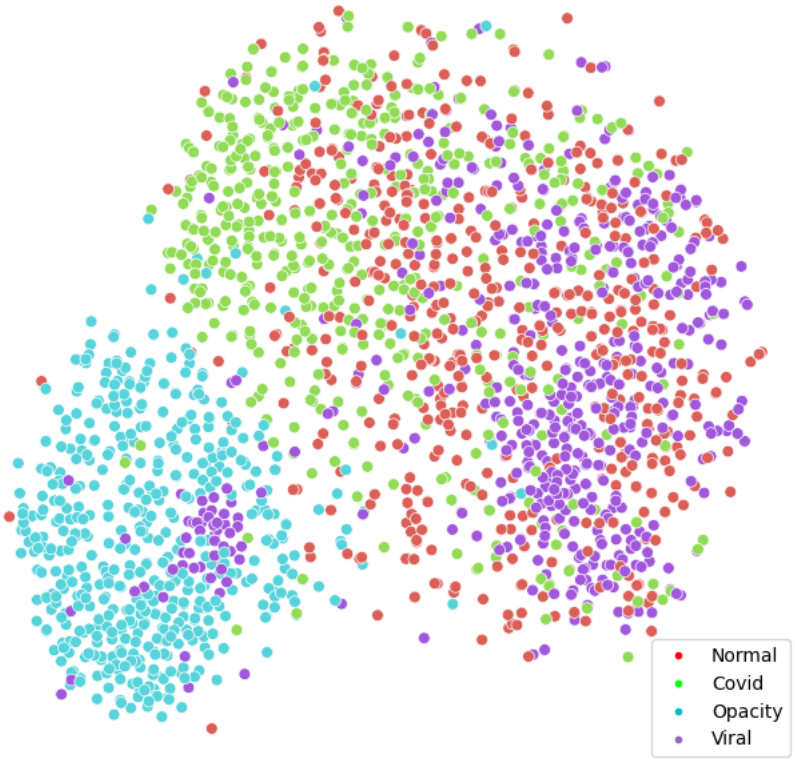


Group 1 Proposal Report Submission

Knowledge-Based Contrastive Learning for Covid-19 Classification


Molly Newquist


Supriya Sundar


Yash-yee Logan


Kiran Kokilepersaud


Adrien Poujon

Georgia Institute of Technology
Electrical and Computer Engineering
ECE 6780: Medical Image Processing

Member	Contribution
Kiran Kokilepersaud	Contrastive learning framework, contrastive learning methodology and analysis, code
Molly Nequist	Contrastive learning analysis and write-up, t-SNE embedding analysis
Adrien Poujon	Clustering methods and analysis, results write-up
Supriya Sundar	Data pre-processing,Clustering methodology and analysis, surveyed existing datasets
Yash-yee Logan	Developed explainability framework, and wrote explainability methodology and analysis, GUI mock-up

Table I: Summary of work and contributions to the report by teammate

KNOWLEDGE-BASED CONTRASTIVE LEARNING FOR COVID-19 CLASSIFICATION

Yash-yee Logan Molly Newquist Kiran Kokilepersaud Supriya Sundar Adrien Poujon

School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

ABSTRACT

In this work we propose a novel supervised contrastive learning framework that integrates radiological observation data with chest x-Ray imagery to improve performance on COVID-19 detection tasks. COVID-19 datasets are smaller and less diverse than previously curated chest x-ray datasets. Additionally, these large datasets usually come with radiological observation data generated during mining of clinical reports for associated information. Contrastive learning is one unsupervised methodology that is able to make use of larger datasets to improve generalization towards smaller datasets, but it doesn't have a way to make use of the radiological information present in the larger dataset as well. By introducing a clustering framework to generate pseudo-labels for the larger dataset, we are better able to choose positive pairs for the contrastive loss and are able to see improved performance on COVID-19 detection tasks after fine-tuning the resulting representation space. Our studies show performance improvements by as much as 2% above state of the art self-supervised and supervised baselines.

I. INTRODUCTION

In December 2019, the novel Coronavirus SARS-CoV-2 was discovered; the disease caused by this virus, known as COVID-19, began a nearly two-year-long pandemic. The late symptom onset of COVID-19 and lack of diagnostic tests available made early detection of SARS-CoV-2 infection challenging, which resulted in widespread infection and subsequent strain on the healthcare system. In order to counter the proliferation of the virus, testing procedures such as the reverse transcription-polymerase chain reaction (RT-PCR) test were developed to diagnose COVID-19 in a clinical setting.

Even with the success of the RT-PCR test, expertise from radiologists is required for the diagnosis and prognosis of COVID-19 related pneumonia. In cases in which patients receive a false-negative from their PCR test, have confounding factors like chronic lung disease, or are from demographics associated with low risk of mortality from COVID-19, it can be challenging for practitioners to catch signs of poor or diminishing lung health. In these cases, information from Chext X-ray scans may be especially helpful. However, interpretation of these scans is time-consuming and is an additional burden on a system that is stretched thin in terms of interpretation demands of radiologists. A study from 2015 [1] showed that radiologists are tasked with interpreting 16.1 images per minute which have contributed to fatigue, burnout, and an increased error-rate. For this reason, investigating approaches that could automatically perform COVID-19 interpretation through machine learning became a major research topic, but the resulting methods were ultimately deemed clinically unreliable [2].

Major tasks for machine learning include training a model to distinguish between different lung conditions with high levels of sensitivity. This effort has proved challenging, especially when an attempt is made to distinguish between COVID-19 related pneumonia and other forms of viral pneumonia, in part due to the similarity of radiological signature between pneumonia classifications. Additionally, conventional deep learning solutions require a large and diverse training pool, which isn't practical when considering the limited amount of available COVID-19 imaging data.

Attempts have been made to solve the problem of limited data access through contrastive learning approaches [3]. Contrastive learning refers to a family of self-supervised methods that make use of pre-text tasks or embedding enforcement losses with the goal of training a model to learn a rich representation space without the need for labels. The general premise is that the model is taught an embedding space where similar pairs of images project closer together and dissimilar pairs of images are projected apart. A classifier can then be trained on top of these learned representations while requiring much fewer labels for satisfactory performance. In this way, contrastive learning presents a way to utilize a large amount of unlabeled data for performance improvements on a small amount of labeled data. In the case of the COVID-19 paradigm, this presents a way to learn useful representations from a large and diverse chest x-ray dataset such as [4] in order to improve performance on the smaller amount of COVID-19 data, even when labels on the two datasets don't match.

Despite the success of these approaches, their usage is complicated by a dependence on finding good augmentations to present the model with similar and dissimilar pairs of images. The process of choosing the best augmentation for a given context remains ad-hoc in the sense that it isn't clear which augmentations are the most suitable. This problem is even more apparent from a medical imaging point of

Terminology	
Word	Meaning
Attrbute Clustering	This is the name we give to our novel methodology (see section III and Fig. 3 for details).
Attributes	This is the name given to the labels aassociated to whole images in the Chexpert dataset. There are 14 of them for each image (see section III paragraph III-A for details).

Table II: Terminology.

view. For example, the disease class may depend on a small localized region within the image, instead of being readily apparent throughout. Consequently, naive choices of augmentations or pre-text tasks can cause occlusion or distortion of regions that are most important to identifying important indicators of a disease.

A more intuitive approach to choosing these positive pairs of images would be on the basis of explicit radiological observations that act as identifying information within Chest X-ray scans. In this work, we introduce a novel supervised contrastive learning strategy that makes explicit use of these observations present within a larger Chest X-ray dataset for performance improvements on COVID-19 detection tasks.

I-A. Our Contributions

Our targeted contributions are:

1) We propose a novel clustering approach on imaging attributes for Chest X-rays to generate labels for a supervised contrastive loss.
2) We propose a novel positive and negative pair selection strategy for contrastive learning in Chest X-ray Images.
3) We show how the proposed strategy can improve performance for Covid-19 classification as well as the generalizability across different datasets, hyperparameters, and data availability settings.
4) We verify the interpretability of our methods through GradCAM [5] and contrastiveCAM [6] on COVID-19 imaging data.

II. RELATED WORKS

II-A. Machine Learning for COVID-19

During the COVID-19 pandemic, a myriad of machine learning papers tried to tackle COVID-19 learning tasks. Here, we present a short review of papers, constrained to applications in radiological chest images. Only a few papers use traditional machine learning; most use deep learning and perform diagnosis or prognosis on CT or Chest X-Ray scans [2]. Papers typically use well-known network architectures such as ResNet-50 or Xception [7], [8] with no modifications. The main preoccupation is classifying scans according to three classes: COVID-19, non-COVID-19 pneumonia and normal. Very few papers distinguish between non-COVID-19, viral, and bacterial pneumonia. Work done in [9] uses an approximate Bayesian CNN to build uncertainty-aware models and distinguish between four classes: normal, COVID-19 pneumonia, viral pneumonia and bacterial pneumonia.

As stated above, deep learning models are the most common way to address COVID-19 detection. Deep Learning is used to extract features from scans allowing a model to discriminate between COVID-19 and other pneumonia. Segmentation of the scans appears to be advantageous and is used in almost all cases [2]. The use of deep transfer learning [10] to build and train classifier models on images is also developed in numerous papers, as a strategy to overcome data challenges. Some original approaches can also be found that propose new classification models, such as Random Forest algorithm based classification models [11].

However, from all these papers, one can say that most are lacking a sufficient external validation that would ground their method for clinical uses [2]. Our work differs from these in that we seek to train a model in a way that exploits additional clinical information through supervised contrastive learning, thereby grounding the methodology with medical intuition.

II-B. Contrastive Learning

Contrastive learning is an approach in which common features of a set pull together the elements of the set while pushing away the elements of another set. These two sets are respectively called positive and negative [12]. This method has been used extensively for medical imaging during the past few years, including for chest X-Ray analysis [13], [14]. Some of the more promising approaches developed in recent years are Momentum Contrast (MoCo) [15], [16], SimCLR [17], and ConVIRT [18].

II-B1. SimCLR

SimCLR features among the most promising works on contrastive learning. It is a self-supervised approach that uses unlabeled data, an approach which is very useful considering that Chest X-Rays typically lack labels. SimCLR uses data augmentation strategies such as crop, color augmentation and blurring, as well as a projection head in the form of a non-linear MLP. With unsupervised images, this network is able to match some supervised pretrained ResNet-50 model [19].

Note, however, that this work is done on natural images and not on X-Rays scans, which are only in black and white and induce more challenges for data augmentation.

II-B2. MoCo

MoCo is another self-supervised approach that uses unlabeled data. It has been adapted for Chest X-Ray data via modifications to the data augmentation methods used in the natural image classification tasks from which this method originates [15]. The contrastive losses in the MoCo framework are used to create dynamic dictionaries, and the encoder within the contrastive learning pipeline associates keys to the data it receives as input. The loss then matches similar keys (based on data augmentation) and dissimilar ones. The issue with this approach is that the dictionaries become larger as the computation is performed; MoCo attempts to solve this problem by updating the keys of dictionaries using a momentum update that takes into account a linear combination of the current query step and the previous one. Additionally, it maintains a last-in-first-out queue structure for keys such that the latest mini-batch of data is queued and the back end of the queue, which is older and outdated, is removed. The advantage of MoCo thus fully relies on this alternative way of update the dictionaries' keys. In MoCo V2, this MoCo approach is integrated on top of the SimCLR methodology [20].

II-B3. ConVIRT

Finally, ConVIRT [18] proposes contrastive learning of radiographical chest images paired with textual reports. It uses the image representations and the paired text, which is agnostic to the medical specialty to ensure easy transfer learning, to perform contrasting between pneumonia chest radiographs and normal ones. The difference between this method and our proposed method is that they use a dual-stream encoder architecture to fuse the representations spaces of image and text-based data. In our work, we use text-based data (radiological observations) to generate cluster labels using a contrastive loss function and then later use that learned representation to classify Covid-19.

There have also been studies on contrastive learning directly applied to the classification of Covid-19 [16], [21], [22]. In these works, the networks were designed for natural images and extended to chest x-ray or computed tomography (CT) images. The challenge here is that natural images contain high diversity in pixel content compared to the above mentioned imagery. Therefore, the augmentations schemes used to generate positive and negative examples in the pretext tasks are non-ideal. Also, despite these works addressing this topic, contrastive learning remains under-utilized and under-valued [19]. Our work differs from typical contrastive learning in the sense that we introduce a novel way to select positive and negative instances in order to introduce a semi-supervised way to do detection of Covid-19. Table III shows a synopsis of what is lacking in existing methods and how our framework will contribute to the field.

II-C. Explainability

Recent questions in the field of artificial intelligence involve the issue of explainability which have been preoccupying researchers, clinicians and philosophers all-together [23]. This approach is crucial to enable the use of artificial intelligence in practical cases. In the domain of biomedical applications, this issue is particularly prominent as artificial systems are used in order to take clinical decisions which involve the health of patients. During the COVID-19 pandemic, a proliferation of novel algorithms and strategies to tackle related challenges had been proposed. However, only a few amongst them achieved in-practice use by clinicians [24].

One can identify at least three ways to perform explainable artificial intelligence in medical imaging: feature analysis, influential region identification and importance of image features [25].

II-C1. Feature Analysis

Feature analysis usually requires a first step of dimensionality reduction which is classically done using principal component analysis (PCA) or t-distributed stochastic neighbor embedding (t-SNE). Another possibility is to use uniformn manifold approximation and projection (UMAP) [26] which is an approach that has been developed in 2020 to perform dimension reduction. UMAP has been used since then for medical applications, notably in 2021 to monitor COVID-19 for a monitoring of COVID-19 disease progression by reducing the dimension of the output of a CNN extracting features from CXRs [27].

II-C2. Influential Region Identification

Identifying influential regions comes down to determining which regions of an image are of interest. Diverse techniques have been used in the past decades but most of them have high computational cost. More recent works are relying on the use of region proposal (deep) networks. Heat-maps are an alternative to region proposals that provide better interpretability [25]. Most recent publications use techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) [5] and contrastive explanations (ContrastCAM) [6]. For instance, [28] uses a probabilistic Grad-CAM saliency map visualization that generates the saliency map taking into account the probability of classes.

Our work takes advantage of these explainability techniques to validate our framework through visual explanations.

III. METHODOLOGY

III-A. Datasets: X-ray

The four datasets of interest to us are the Covid-19 Radiography database [29], [30], the COVID Qu Dataset [31], the Covid X Dataset [32] and the Chexpert dataset [4]. The Covid-19 radiography database, COVID Qu, and COVID X datasets are the ones that we use to evaluate

Algorithm	Uses Larger Dataset	Medically Intuitive	Wider Pool of Positive Pairs
Supervised Learning	✗	✗	✗
Unsupervised CL: SimCLR	✓	✗	✗
Unsupervised CL: Moco V2	✓	✗	✗
Supervised CL (Ours)	✓	✓	✓

Table III: A synopsis of the weaknesses of existing methods how our proposed method counters these limitations.

COVID-19 detection performance on. The Chexpert dataset is used only for our contrastive learning experiments in order to develop a good representation space on which COVID-19 data can be fine-tuned on top of.

The Chexpert dataset has a training set consisting of approximately 224,316 Chext X-Ray scans across 65,240 different patients. Images start at a resolution of 312 X 312. For each X-Ray scan in this dataset, there exists a vector of uncertainty labels that indicates the presence or absence of 14 different radiological observations. These observations are:

- No finding
- Enlarged Cardiom
- Cardiomegaly
- Lung Lesion
- Lung Opacity
- Edema
- Consolidation
- Pneumonia
- Atelectasis
- Pneumothorax
- Pleural Effusion
- Pleural Other
- Fracture
- Support Devices

The Covid-19 Radiography dataset contains 21,116 Chext X-Ray scans of size 299x299 across 4 classes: Covid, healthy, viral pneumonia, and lung opacity. No explicit train and test split was provided for this dataset, so we chose to split it by sampling 500 images from each class to create a balanced test set containing 2000 images. In order to ensure no sources of bias, the sampling strategy was additionally constrained such that the test set would not contain any images belonging to a patient that also existed in the training set. No explicit information regarding the patient data was provided. The authors of this paper were contacted and it was found they used a sampling strategy that got a single scan from each patient. Furthermore, to get conclusive results on this dataset, we performed a 3-fold cross-validation on different subsets of patient splits to account for any bias regarding specific distributions of patients.

The COVID-QU dataset contains 33,920 images from three classes: Normal, non-COVID, and COVID-19. The authors provided a balanced test set containing 2000 images from each class. No explicit information regarding the patient makeup was reported, but the authors ensured that images in the test set did not correspond to images from patients in the training set. This dataset also has infection segmentation masks of the regions that have COVID-19. For this reason, this dataset was utilized for classification experiments as well as experiments that verified the interpretability of our models.

The COVID X dataset contains 30,000 images from two classes: COVID-19 and non-COVID-19. The dataset is drawn from a multi-national cohort of 16,600 patients. The test set provided by the authors is separated by patient identity from the training set and contains 200 images drawn from each of these two classes to ensure a balanced test set. In this dataset, they also report the source of the data and split the training and test set based on dataset source as well as patient.

III-B. Data Pre-Processing

Prior to actual utilization by the models, it is crucial to pre-process this data tailored to suit the specifications of each framework.

Although chest radiographs are considerably advantageous in analyzing the lung pathology, the low contrast and varied sizes of real-time images collected, the potential noise it could contain, and possible duplicates would add a challenge in our prediction. We checked the dataset for missing or null images and there were none. Thus, we did not perform any kind of interpolation or over and under sampling.

Data augmentation was implemented in the pre-processing phase to mitigate overfitting and make the model adaptable to generic data. Since data acquisition methods were varied from the different datasets utilized, we reform the size of all images to 224 pixels in grayscale channel to overcome data variability. Thus, transformations such as image resizing, horizontal flip, color jitter and conversion to grayscale were performed. Data normalization was performed with mean and unit standard deviation.

III-C. Supervised Contrastive Learning for COVID-19 Classification

The idea behind contrastive learning is to learn a better representation space to improve performance on downstream tasks. That is, by using a contrastive loss over a traditional cross-entropy loss, we can find a representation space that learns discriminating features in our images before trying to explicitly classify for class labels. This is particularly advantageous within the medical domain because features

that differentiate between disease classes are often highly localized. For this reason, there is considerable overlap between classes and it becomes a difficult task to learn discriminating features.

In self-supervised techniques like SimCLR and Moco, the representation space is created by maximizing the similarity between an image and a single augmentation of that image. Each image in a batch of images $\{x_i\}_{i=1}^{N}$ is augmented to create a set $\{\tilde{x}_i\}_{k=1}^{2N}$ of original and augmented images. Each image with index $i \in I = \{1 \dots 2N\}$ is passed through an encoder network $f(\cdot)$ and then compressed through a projection head $G(.)$, which is discarded at the end of training. Let z_i be the output of the projection head. Then the typical contrastive loss is given by

$$L_{self} = \sum_{i \in I} \log \frac{exp(z_i \cdot z_j / \tau)}{\sum_{a \in A(i)} exp(z_i \cdot z_a / \tau)} \quad (1)$$

where $A(i) = I \setminus i$ is the set of the single positive and all negative instances for image x_i and τ is a temperature scaling parameter.

While this approach may teach the model more about the image distribution generally, it is not tailored to any specific domain or task. Supervised contrastive learning [33] addresses this by incorporating label information during the creation of positive and negative sets. If we allow multiple positive instances, specifically such that all images and augmentations with the same label are considered positive instances of each-other, we can learn a representation space better suited to classification tasks. To accommodate, the loss function becomes

$$L_{sup} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} exp(z_i \cdot z_a / \tau)} \quad (2)$$

where $P(i) \subset A(i)$ is the set of indices for all positive instances for image x_i. Note that the question of how to construct the set $P(i)$ is open ended. Typically, this has been done using the same image labels used in the downstream classification task. The major innovation of our work is a method by which to construct $P(i)$ within a transfer learning pipeline and in a way that incorporates medical intuition.

III-D. Transfer Learning Pipeline

Even though there are thousands of datasets available currently on the novel SARS-CoV-2 virus, most of this data is not effective in utilization in a classification model due to several factors. There are significant challenges in data collection and access due to privacy concerns and the closed-source nature of patient data. There is also lack of aggregation among the datasets available globally and it is unlabeled for the larger part. A large portion of the datasets have been collected from a specific geographic location which has resulted in selection bias when applied in other countries. Images from children have been used for "Non-COVID" category and images from adults have been used for "COVID" category in many models, which has invariably caused bias. To overcome these problems, we propose the two-stage transfer learning framework introduced in Figure 3. The first stage is to use supervised contrastive learning to pretrain an encoder network; here the criteria for positive pair selection is based on the 14 radiological observations in the Chexpert dataset. By pre-training the representation space based on clinical observations that are highly relevant to COVID-19 diagnosis, we are able to incorporate knowledge from larger pre-pandemic chest X-ray datasets into our COVID-19 classification problem.

To do this, we follow a similar strategy for supervised contrastive learning as [33]. We chose the encoder $f(\cdot)$ to be ResNet-18 [35], whose output is a 512×1 dimensional vector r_i. ResNet-18 was used for its smaller size and convenience use in both supervised and self-supervised methodologies; this allows easier comparison of our method to other techniques. The vector r_i is further compressed through the projection head $G(.)$, which we set to be an multi-layer perceptron with a single hidden layer. The output of $G(.)$ is a 128×1 dimensional embedding z_i.

Any of the 14 radiological observations in the Chexpert dataset could used as a label to determine the positive and negative pairs of images. A naive approach is to use the presence of a single radiological attribute as a label for supervised contrastive learning; for example, one might collect all images with lung lesions present into the positive example set. However, this method ignores the fact that the radiological attributes are not mutually exclusive. A single image might be labeled as positive for pneumonia and pleural effusion, and it may be the combined presence of these conditions that is the best indicator for COVID-19, for example. Answering the question of which observations or combination of observations leads to the best trained representation space can help identify which characteristics of COVID-19 distinguish it from other disease states.

In stage 2 in Figure 3, we take the previously trained encoder and freeze its weights. A linear layer is attached to the output of this encoder, and images from the Covid-19 dataset are passed through this network to produce a prediction $\hat{y}$. A cross-entropy loss between this prediction and the ground truth label y is back-propagated to train the linear classifier, which fine-tunes the representation learned in step 1 to the task of COVID-19 pneumonia detection. In this way, we are able to use an encoder trained with a large amount of bio-marker data to inform the detection task of the labels in the much smaller Covid-19 dataset.

III-E. Clustering on Radiological Observation Labels

As discussed, there is strong motivation for incorporating multiple radiological observations into the positive examples selection criteria for stage 1. To this end, we developed a

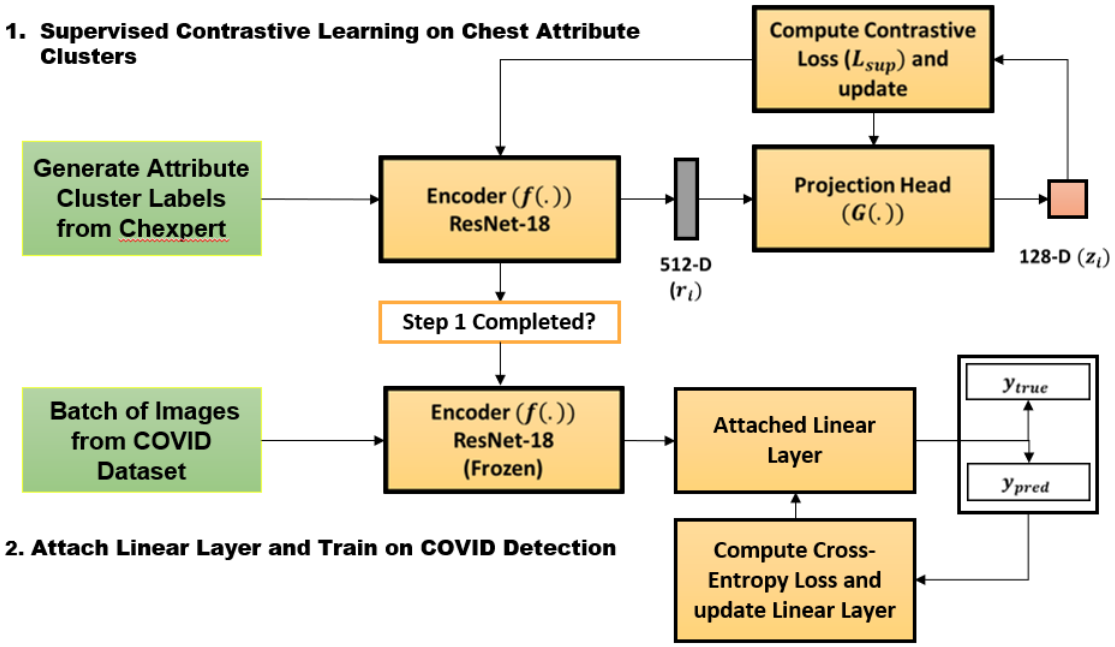


Fig. 2: Overview of supervised contrastive learning and linear fine-tuning steps. 1) Supervised Contrastive Loss on attribute labels from large Chexpert dataset. 2) Attach linear layer and fine-tune representations on smaller Covid-19 detection data.

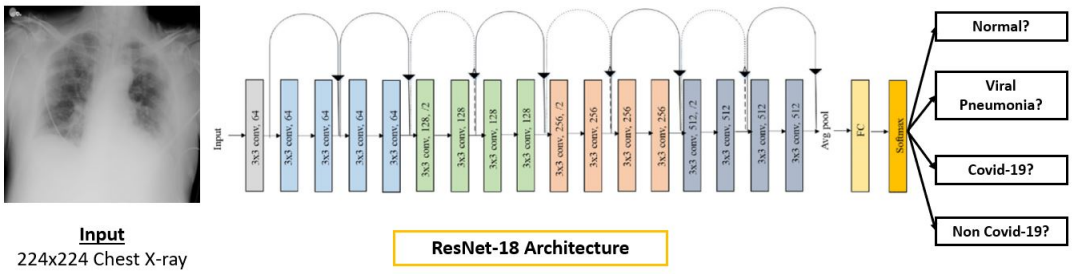


Fig. 3: Overview of Resnet-18 [34] architecture and inputs for classification on Covid Kaggle Dataset.

unique approach to generate pseudo-labels for supervised contrastive loss based on clustering. Recall that each of the 226,000 Chexpert images is labeled with 14 different binary observations. Using these observations, we generate a $226,000 \times 14$ image characterization matrix. With a cluster size of 200, determined via trial and error, we perform K-Means clustering on the characterization matrix and generate new labels based on the clustering. This gives us the labels needed to identify the positive pairs needed for contrasting learning while allowing us to incorporate more than one radiological observation in that label.

In addition to the generation of appropriate labels, clustering techniques are also utilized in accurate identification of observations that distinguish one attribute from the other. In addition to the K-Means, other clustering algorithms such as Gaussian Mixture and Mini-Batch K-Means were tested. The relevance of specific attributes with regard to

Fig. 4: Cluster analysis for identification of relevance between attributes.

one another could be clearly visualized with the Gaussian Mixture Clustering. Specific instances of cluster from the final cluster size of two hundred were visualized uniformly and it could be observed that different attribute clusters were

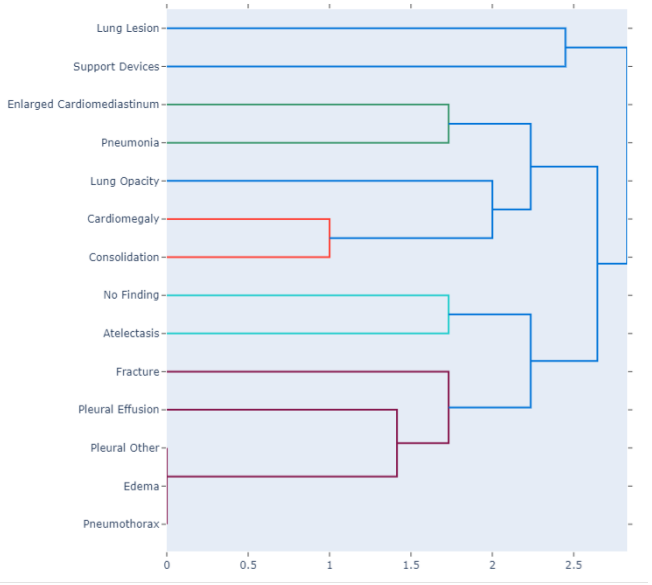


Fig. 5: Hierarchical dendrogram showcasing the relationship between clusters.

in closer proximity in each instance. This radar plotting helped us comprehend the flow and gradual aggregation of clusters of each biological attribute as shown in Fig. 4. It is seen that while lung opacity and plural effusion are similar in the twelfth cluster, lung opacity and pneumothorax are more relevant to each other in the sixteenth cluster. Thus, this clustering helps us determine the stage at which observations that distinguish each attribute can be identified.

A dendrogram depicting the hierarchy of clusters between the fourteen attributes was plotted as well. The distance at which two distinct clusters combine and the number of clusters appropriate to the dataset was analyzed. By employing complete linkage type hierarchical clustering, it was observed that lung lesion and support devices clusters were similar to each other initially as was enlarged cardiomediastinum and pneumonia to each other. The relevance of the similarity of clusters and their aggregation stages over time is depicted in Fig. 5 .

III-F. GradCAM and Contrastive Explanations

In order to understand the rationale behind our model's predictions, we use two frameworks to provide visual explanations for the model's decisions. The first is GradCAM [5], which uses the gradients of a specific label within a classification network at the last convolutional layer to create a visualization highlighting regions that were pertinent to the model making its prediction. It provides visual answers to the question "Why P?", where P represents the predicted class. GradCAM generates a localization map of width u and height v for any class c. The first step in computing this map involves computing the gradient of the predicted class score with respect to the feature activation of the final convolutional layer $\frac{\delta y^c}{\delta A^k}$. Equation 3 shows how a global average is then taken of these gradients across the height and width dimensions to compute the importance of each weight α_k^c.

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\delta y^c}{\delta A_{i,j}^k} \tag{3}$$

Next, the importance of positive feature maps k for a predicted class c is expressed with an inner product, as shown in Equation 4. This localization map is called GradCAM $L_{GradCAM}^c$.

$$L_{GradCAM}^c = ReLU(\sum_k \alpha_k^c A^k) \tag{4}$$

The second framework we use to generate visual arguments to justify network predictions is known as Contrastive-CAM [6]. These explanations are constrained based on an additional class, Q, and they visually answer questions of the form, "Why P, rather than Q?". Contrastive explanations are embedded into the existing, "Why P?" explanations that GradCAM provides. Specifically, contrastive explanations uses the networks discriminative knowledge, stored in its weight W and bias b terms, to separate between the prediction of an input as P and the prediction of an input as Q. In terms of the network's representation space, contrast is the difference between manifolds that predict an input as P and as Q. Gradients are used to measure the change required to generate a contrastive manifold from a learned manifold.

Contrast for class Q when an image is predicted as class P is computed by back-propagating the loss $J(P, Q, \theta)$ between P and Q as shown in Equation 5. θ are the network's parameters.

$$Contrast_Q = \frac{\delta J(P, Q, \theta)}{\delta \theta} \tag{5}$$

When a network makes a prediction P, the gradients obtained from Equation 5 represent, "Why P, rather than Q?" and can be plugged into a GradCAM framework to showcase contrastive explanations.

III-G. Targeted Performance and State of the Art

Since new images are constantly being added to the Covid Chest X-ray dataset, there are no standard performance metrics on this dataset that we can seek to meet. For example, the original version of the dataset that appeared in the paper [30] was approximately 7 times smaller than the current one, with much less diversity in terms of patient distributions. For this reason, the authors were able to achieve nearly perfect performance values on the order of nearly 99% accuracy, specificity, and sensitivity. Considering this, our goal is not to out-perform a specific target value; instead, we seek to out-perform specific methods that we can implement

and run on current data with a cross-validation procedure. These methods are traditional supervised learning with a standard cross-entropy loss as well as two state-of-the-art self-supervised methodologies [17], [20]. We will measure performance in terms of accuracy, sensitivity, specificity, and f1-score averaged across different folds in a 5-fold cross-validation procedure on the Covid dataset of interest to us. This will provide a comparison between our methodology and other methods as well as present a better way to pre-train a network for Covid-19 classification. The above mentioned metrics are computes as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{9}$$

where TP = true positive, TN = true negative, FP = false positive and FN = false negative

IV. COMPARATIVE RESULTS

IV-A. Test performances on three datasets, against three others methodologies

We dedicate this section to the presentation of some of the results achieved by our framework. To begin with, we compare usual metrics for test performance of our approach versus three other approaches: a classical supervised approach, the state of the art MoCo v2 and SimCLR. These results are presented on Table IV which thus constitute a comparison with other methodologies for a given dataset on the basis of local runs of these methods. Each result is an of three random initializations and three folds of the datasets, and we present them for two different datasets: Covid Kaggle and Covid_X. For each metric, our attribute clustering method achieves better results based on local runs of the methods. The largest increase is found for precision. These results are very encouraging for eventual application in medical context as they allow overall better confidence in the diagnosis made by the model.

To further our comparison with other methods, We also visualise the performance of each approaches through confusion matrix om Fig. 6. These matrix show how well these methods do for distinguishing different classes. It is very clear that the fully supervised and SIMCLR approaches performs really poorly at differentiating between COVID-19 and non-COVID-19 or normal radiography. Similarly, our Attribute clustering method and the MoCo V2 algorithm have difficulties and still confuse normal and non-COVID-19 scans with COVID-19 one. And yet, these to methods nonetheless manage to get better results as the true positive ratio is larger by two order of magnitude. Even if it is by a small margin comparing to MoCo V2, our Attribute clustering displays better metrics within the confusion matrix, once again indicating better applicability in medical context after further improvements.

As explained in previous sections, the novelty of our method from other contrastive learning approaches by using clusters as a way to create image pair sets instead of using image augmentation. Thus, one of the key components of our approach is the way we choose to build our clusters. Instead of using clustering algorithms such as KNN or GMM over the whole attributes of Chexpert dataset, one might as well directly perform contrastive learning using each attribute individually. We illustrate the results of such an approach on Table V. When compared to the accuracy obtained in Table IV, it is seen that the performances are a lot worst as accuracy is almost divided by a factor 2. This shows that the representation spaces resulting from a single-attribute training approach are less adequate for characterizing COVID-19.

The predictability of the four methods are shown on Fig. 7. The graph represents how our model generalizes from a training setup to a testing environment. At best it is expected that if the training were perfect it would allow results for the testing being equally high at those for the training. In practice, as the training phase is done over a specific subset of the data, the results obtained in the testing phase should slightly diverge from the identity curve. Our Attribute clustering method is closer to the identity curve for the Accuracy, F1-score and Precision metrics. However, it is clear that it exhibits a poor performance for the precision test on this analysis when compared to the results of other methods.

IV-B. Robustness of the method compared with robustness of three other methodologies

We now present some experiments to ensure the robustness of our models to variations of the parameters. A first parameter to look at is the size of the training sets. We perform an analysis of the average accuracy over three trials obtained after having trained on a fourth, a half, three fourth or the full training set on Fig. 9 and Fig. 8. The former was performed using the COVID Kaggle dataset and shows how our model out performs the others for every training set sized considered. The later is realized after results on the QU dataset, it shows similar results although the difference of accuracy for the three contrastive learning approaches are way lower. In both cases the robustness of our attribute clustering method to the size of training set seems very strong.

Especially, our methodology compare very well with other methods on Fig. 9 and Fig. 8. We can see that for a given sample size of the dataset our clustering attribute

Test Performance on Covid Kaggle Dataset				
Method	Accuracy	Precision	Recall	F1-Score
Supervised	0.6950	0.7787	0.6984	0.6964
Moco v2	0.6910	0.7752	0.6953	0.6782
SimCLR	0.6789	0.7734	0.6800	0.6562
Attribute Clustering	0.7284	0.8194	0.7144	0.7238
Test Performance on Covid_X Dataset				
Method	Accuracy	Precision	Recall	F1-Score
Supervised	0.6683	0.7570	0.4182	0.5677
Moco v2	0.7168	0.7620	0.448	0.5950
SimCLR	0.6991	0.7820	0.4282	0.5800
Attribute Clustering	0.7184	0.7830	0.4346	0.5835

Table IV: Test performances on different datasets and for three methods comparing to our attribute clustering approach.

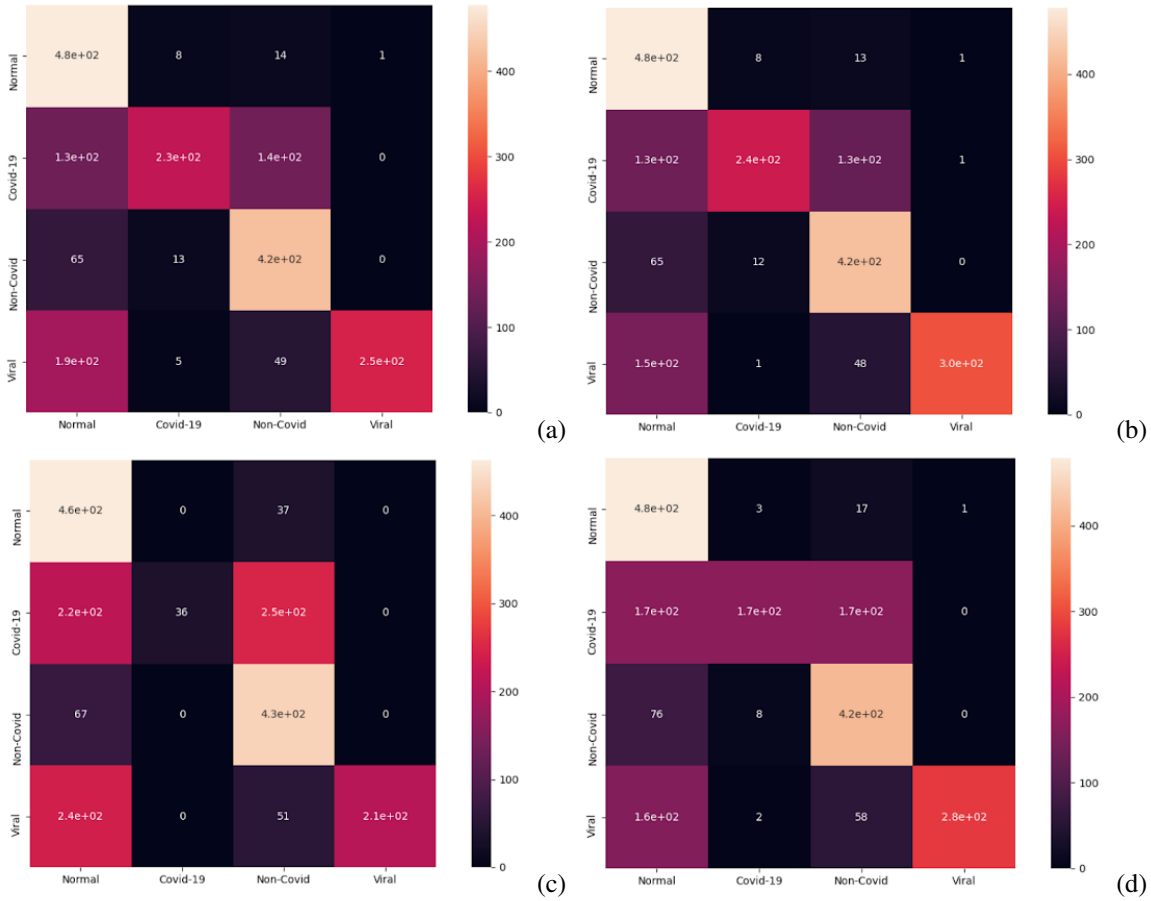


Fig. 6: Confusion matrix for the MoCo V2 (a), Attribute clustering (b), Fully supervised (c) and SimCLR (d) methods.

method stays on top of the accuracy values. Note that the accuracies for Moco v2, SimCLR and Supervised learning were obtained through local runs of these methods.

To further the testing of our model's robustness, we also tried to evaluate the influence of the random initialization on its test accuracy comparatively to the fully supervised, SimCLR and MoCo v2 methods. The results of this analysis are shown on Table VI and illustrated on Fig. 10. It is very clear that the attribute clustering method gets better accuracy no matter matter the random seed being used and this for almost all percentage of the training data being used. Only the fully supervised approach can sometimes achieve better

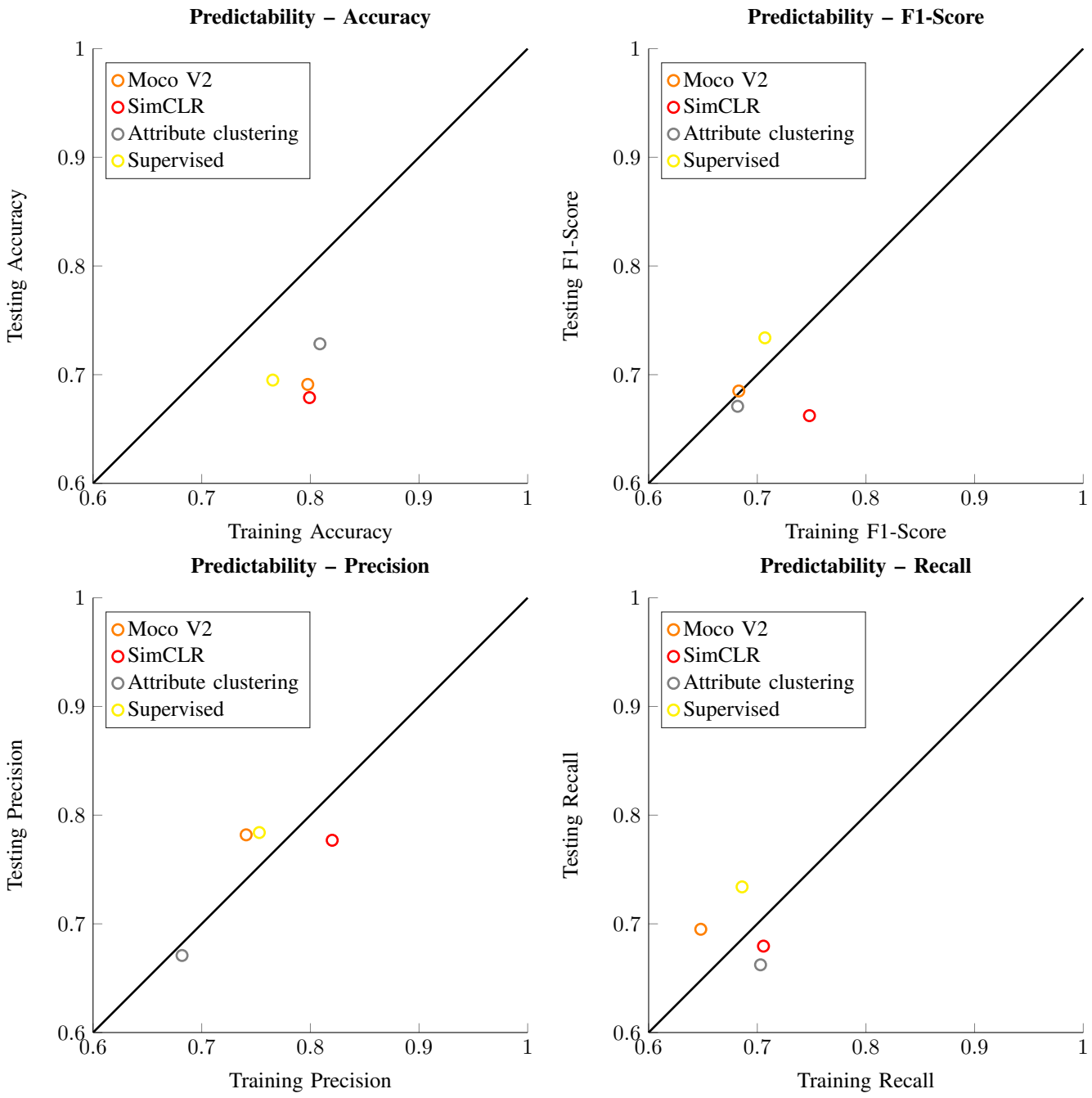


Fig. 7: Average Accuracy, F1-Score, Precision and Recall for training of different subset size of the training set for Kaggle dataset. The black line represents the identity curve.

accuracy on Fig. 10, and yet this comes at the expense of high variability and very low robustness to the random seed for this approach. Overall, our approach has the highest mean accuracy and one of the lowest sensibility to the random initialisation, with MoCo V2 being the only one with a lower standard deviation.

Finally, we present a sensitivity map illustrating the influence of the learning rate coupled with the considered batch size on the accuracy om Fig. 11. This heatmap displays an area of high accuracy values for a batch size around 64 and a learning rate around 0.0005. Choosing our hyperparameters in that range provides additional robustness

Test Performance on Covid Kaggle Dataset	
Attribute	Performance (Accuracy)
Atelectasis	42.90%
Cardiomegaly	41.60%
Consolidation	47.35%
Edema	48.90%
Enlarged Cardiomedia	41.60%
Lesion	44.80%
Pleural Other	45.90%
Pneumonia	47.20%
Pneumothorax	39.30%
Lung Opacity	53.21%

Table V: The contrastive learning step is trained with each attribute individually. Then Covid-19 Classification is performed. The resulting representation spaces were all less adequate for characterizing Covid-19.

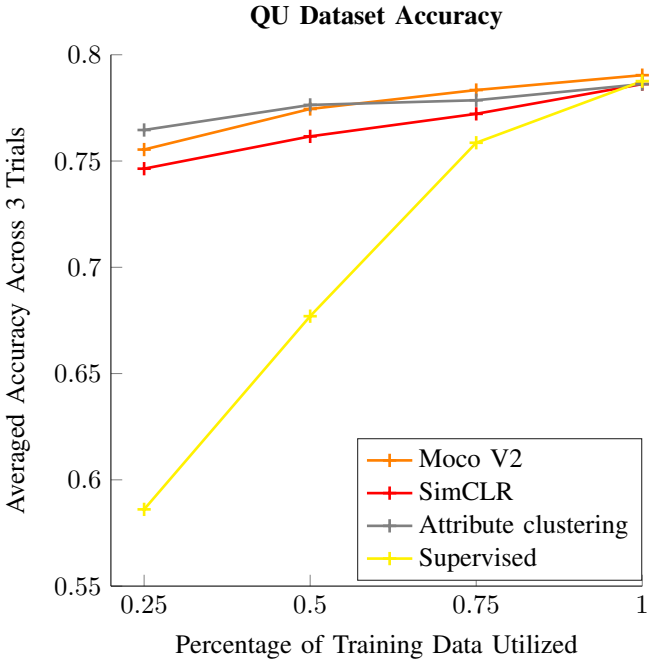


Fig. 8: Average accuracy for training of different subset size of the training set for Kaggle dataset. Our method achieves better results no matter the considered size.

to our method when applied to other datasets by ensuring that small variations would still stay on the plateau characterized by high accuracy values.

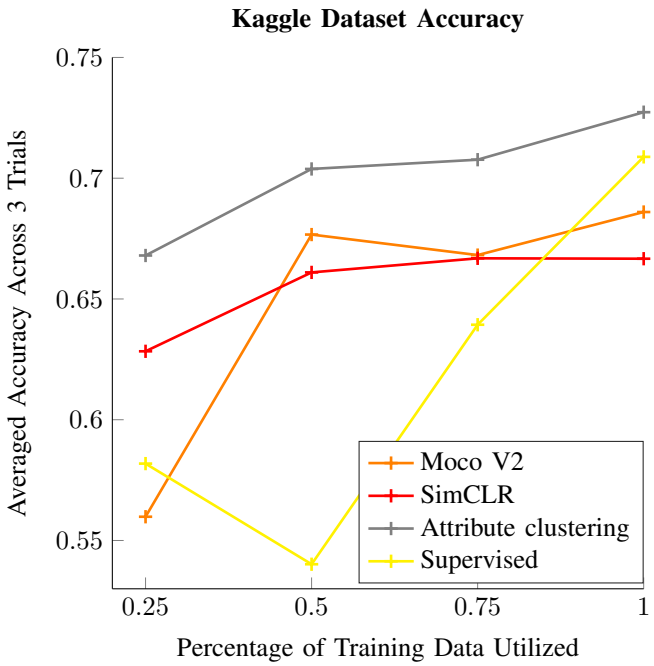


Fig. 9: Average accuracy for training of different subset size of the training set for COVID-19 QU dataset. Results for other methods are better for this dataset then the Kaggle dataset, and yet our method still remains one of the best.

Test Performance on Covid Kaggle Dataset		
Method	Mean Accuracy	STD
Supervised	0.6910	.0307
SimCLR	0.6789	.0160
Moco v2	0.6950	.0035
Attribute Clustering	0.7284	.0120

Table VI: Attribute clustering possesses better mean and lower standard deviation then other methods. This ensure the robustness of our approach to the randomness of initialization.

V. DISCUSSION

V-A. Radiological Observation Supervised Contrastive Learning

In order to understand why our method performed better than the chosen baselines, we need to explore literature regarding representation spaces and why contrastive learning works in the first place. The overall goal of any contrastive learning pre-training is to first devise a good representation space on which the task of interest can be trained. This leads to questions as to what defines a good "representation." The authors of [36] identify several factors that are especially relevant to the setting of this paper. The first is spatial or temporal coherence, which means that nearby observations

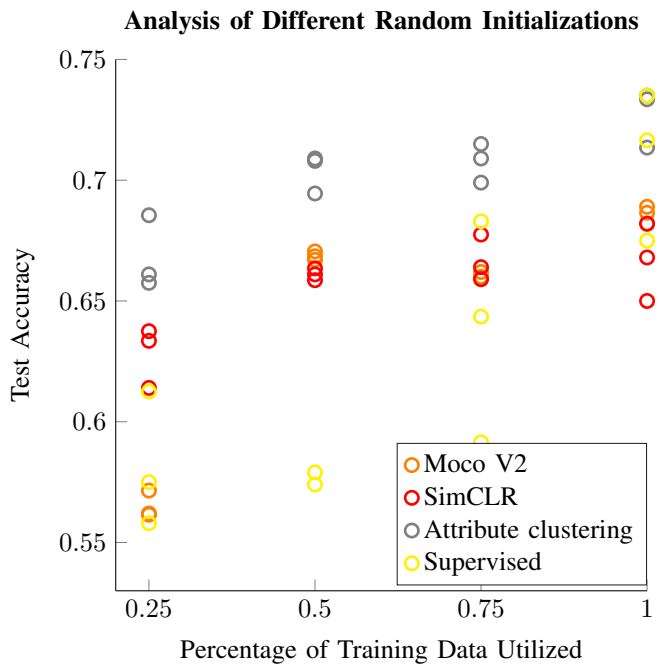


Fig. 10: This graph shows the results obtained for four methods when the initial seed is modified. It illustrates the consistency of our method.

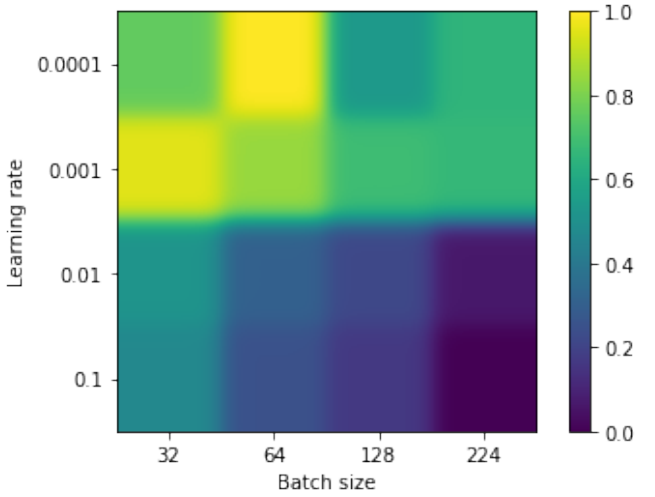


Fig. 11: This heatmap illustrates the dependency of our model's results on the choice of batch size and learning rate. Performance is evaluated is term of accuracy. A plateau of high accuracy values is observe around batch size = 64 and learning rate = 0.0005

are associated with the same category. In this case, we are able to identify samples that should be nearby to each other through the use of labels generated from radiological observation data. It can be argued that images with similar radiological attributes belong to a distinct category which distinguishes them from those with different sets of attributes. Therefore, shaping the representations in a supervised manner pushes objects in the same category closer together in the representation space than if this was done in a fully unsupervised fashion.

Another aspect of good representations is that they should have multiple explanatory factors and a hierarchical organization of explanatory factors. Explanatory factors describe the idea that the underlying distribution should be the result of elements that can be disentangled from each other. A hierarchical organization implies that the concepts utilized in creating the representations should be useful to describing the world around us in the sense that some factors are more and less relevant to the resultant representation space. In the case of our work, we argue that these "factors" are the radiological observation data we are able to integrate into the training process. This data effectively describes the conditions present within each X-ray scan and, by using it as the factors by which to shape the representation of these X-ray scans, we are better able to establish an organization that aligns with real-world considerations. This may explain why our method out-performs completely unsupervised contrastive learning techniques where no usage of medically aligned intuition is used to shape the representations, thus lacking any form of explanatory factors.

Additionally, we can understand the results we observe through the work of [37]. The authors of this paper identify that the gap between contrastive learning and supervised learning increases as the granularity of the task increases. In other words, fully unsupervised approaches do worse as the features that represent the class are more sparsely represented in the data. This is the case in the medical domain because it is oftentimes the case that the regions of most relevance are not present throughout the image. Small localized regions are oftentimes the ones we are most interested in detecting. Contrastive learning performing worse in this case may be the result of standard methods having an over-reliance on augmentations originating from a single image to generate positives. This problem isn't as apparent with our method because we are choosing positives from a larger set of images. Instead of relying on an augmentation to generate the positive set we use medically consistent information to identify images that should exist closer to each other in a representation space. In this way, the model may be better able to overcome this granularity issue with contrastive learning. This is evidenced by our model performing significantly better when the number of classes increased from 2 to 4. With more classes, finer grain differences between pneumonia and COVID-19 based pneumonia are necessary to distinguish which our model is shown to better rectify.

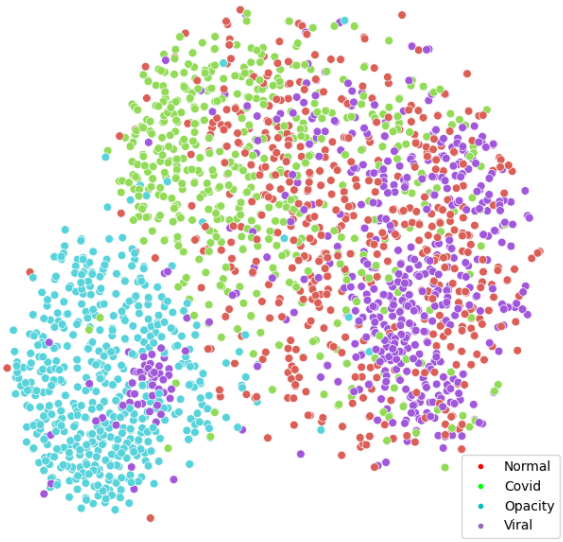
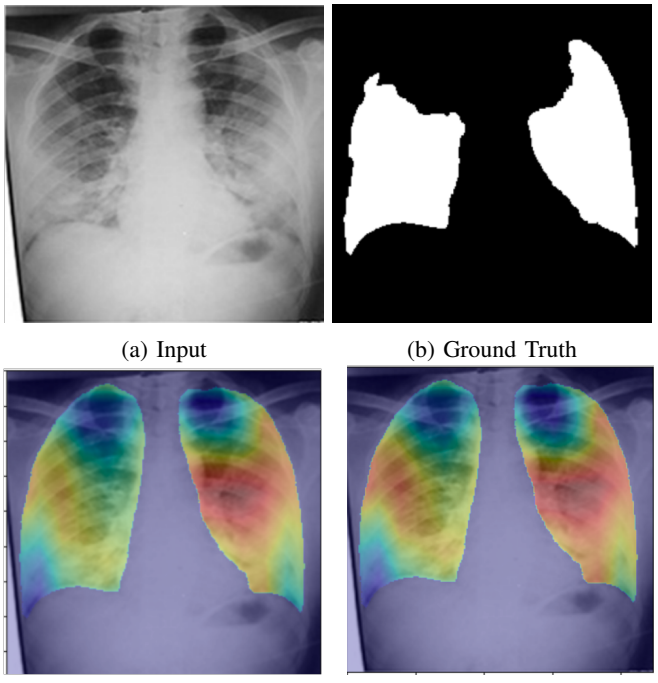


Fig. 12: t-SNE [38] embedding space visualization of different classes within the Covid-19 test set. It can be observed that Covid can be confused with viral pneumonia and regular healthy images.



(a) Input (b) Ground Truth

(c) GradCAM (d) ContrastCAM

Fig. 13: Visual explanations of important (red) regions corresponding to the model's prediction. Subfigures: (a) A Covid Positive X-Ray (b) Localized Covid-19 Infection (c) Visual explanation: Why is this Covid-19? (d) Visual explanation: Why is this Covid-19 rather than Normal?

V-B. GradCAM and Contrastive Explanations

To demonstrate both GradCAM and contrastive explanations we present some visual explanations from each method in Fig. 13. Red and blue regions represent areas that are of most and least importance to the model when it made its predictive decision respectively. For both applications, a VGG-19 model pretrained on ImageNet was used with the final fully connected layer replaced and trained from scratch to learn the three classes contained in the Covid-19 Qu dataset. This dataset contains pixel level labels of regions within the lungs where Covid-19 is present. This pixel level description has no use in our methodology except for interpreting our visualization. It substitutes the eye of a radiologist to properly interpret the scans and allows us to make sense of the areas highlighted by our visual explanation. The model was trained for 25 epochs with a stochastic gradient descent optimizer along with a learning rate of 0.001 and momentum of 0.9. All images were resized to be 224×224 and normalized to have zero mean and unit standard deviation. Random images within the training dataset were horizontally flipped and rotated by 10 degrees.

Figure 13 illustrates some visual heatmaps generated by GradCAM and ContrastCAM. The pixel-level location of Covid-19 ground truths removed the need of a radiologist to confirm the validity of our method. In Fig. 13c and Fig. 13d, the model makes a prediction for Covid-19 based on accurate pixel locations since the red and yellow regions coincide with the ground truth. The blue areas on the heatmap correctly indicate the model detected an absence of Covid-19 features in those regions.

VI. CONCLUSION

Contrastive learning is useful to the problem of COVID-19 detection because it allows us to integrate a large pool of chest x-ray images into the training process. It relies on a strategy to define positive and negative pairs of images. We show in this paper that one such strategy is to utilize radiological observations to generate labels via clustering that identify Chest X-rays with similar structures in common. Training on these labels with a supervised contrastive loss leads to an improved representation space that is able to display better generalization across COVID-19 datasets as well as different hyperparameter and training paradigms compared to state of the art self-supervised and fully supervised methods.

VII. REFERENCES

[1] Robert J McDonald, Kara M Schwartz, Laurence J Eckel, Felix E Diehn, Christopher H Hunt, Brian J Bartholmai, Bradley J Erickson, and David F Kallmes, "The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist

workload," *Academic radiology*, vol. 22, no. 9, pp. 1191–1198, 2015. 2

[2] Michael Roberts and Derek Driggs et al, "Fcommon pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans.," *Nature Machine Intelligence VOL 3 MARCH 199-217*, 2021. 2, 3

[3] Phuc H Le-Khac, Graham Healy, and Alan F Smeaton, "Contrastive representation learning: A framework and review," *IEEE Access*, vol. 8, pp. 193907–193934, 2020. 2

[4] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al., "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, vol. 33, pp. 590–597. 2, 4

[5] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626. 3, 4, 8

[6] Mohit Prabhushankar, Gukyeong Kwon, Dogancan Temel, and Ghassan AlRegib, "Contrastive explanations in neural networks," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 3289–3293. 3, 4, 8

[7] Enzo Tartaglione, Carlo Alberto Barbano, Claudio Berzovini, Marco Calandri, and Marco Grangetto, "Unveiling covid-19 from chest x-ray with deep learning: A hurdles race with small data.," *Int. J. Environ. Res. Public Health 17*, 2020. 3

[8] Zheng Wang, Ying Xiao, Yong Li, Jie Zhang, Fanggen Lu, Muzhou Hou, and Xiaowei Liu, "Automatically discriminating and localizing covid-19 from community-acquired pneumonia on chest x-rays.," *Pattern Recognition 110*, 2021. 3

[9] Biraja Ghoshal and Allan Tucker, "Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection.," *rXiv:2003.10769v2*, 2020. 3

[10] Nandhini Subramanian, Omar Elharrouss, Somaya Al-Maadeed, and Muhammed Chowdhury, "A review of deep learning-based detection methods for covid-19.," *Computers in Biology and Medicine*, 2022. 3

[11] Feng Shi, Liming Xia, and Fei Shan et al, "Large-scale screening to distinguish between covid-19 and community-acquired pneumonia using infection size-aware classification.," *Physics in Medicine Biology*, 2021. 3

[12] Aaron van den Oord and Oriol Vinyals, "Representation learning with contrastive predictive coding.," *arXiv preprint arXiv:1807.03748*, 2018. 3

[13] Anuroop Sriram, Matthew Muckley, Koustuv Sinha, Farah Shamout, Joelle Pineau, Krzysztof J. Geras, Lea Azour, Yindalon Aphinyanaphongs, Nafissa Yakubova, and William Moore, "Covid-19 prognosis via self-supervised representation learning and multi-image prediction.," *arXiv preprint arXiv:2101.04909v2*, 2021. 3

[14] Yen Nhi Truong Vu, Richard Wang, Niranjan Balachandar, Can Liu, Andrew Y. Ng, and Pranav Rajpurkar, "Medaug: Contrastive learning leveraging patient metadata improves representations for chest x-ray interpretation.," *arXiv preprint arXiv:2102.10663v1*, 2021. 3

[15] Hari Sowrirajan, Jingbo Yang, Andrew Y. Ng, and Pranav Rajpurkar, "Moco-cxr: Moco pretraining improves representation and transferability of chest x-ray models.," *arXiv preprint arXiv:2010.05352v3*, 2021. 3, 4

[16] Xiaocong Chen, Lina Yao, Tao Zhou, Jinming Dong, and Yu Zhang, "Momentum contrastive learning for few-shot covid-19 diagnosis from chest ct images," *Pattern recognition*, vol. 113, pp. 107826, 2021. 3, 4

[17] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607. 3, 9

[18] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz, "Contrastive learning of medical visual representations from paired images and text .," *arXiv preprint arXiv:2010.00747v1*, 2020. 3, 4

[19] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations.," *Proceedings of the 37 th International Conference on Machine Learning*, 2020. 3, 4

[20] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020. 4, 9

[21] Jinpeng Li, Gangming Zhao, Yaling Tao, Penghua Zhai, Hao Chen, Huiguang He, and Ting Cai, "Multi-task contrastive learning for automatic ct and x-ray diagnosis of covid-19," *Pattern Recognition*, vol. 114, pp. 107848, 2021. 4

[22] Tingyi Wanyan, Hossein Honarvar, Suraj K Jaladanki, Chengxi Zang, Nidhi Naik, Sulaiman Somani, Jessica K De Freitas, Ishan Paranjpe, Akhil Vaid, Jing Zhang, et al., "Contrastive learning improves critical event prediction in covid-19 patients," *Patterns*, vol. 2, no. 12, pp. 100389, 2021. 4

[23] Andrés Páez, "The pragmatic turn in explainable

artifcial intelligence (xai).," *Minds and Machines, 29:441–459*, 2019. 4

[24] Felipe Giuste, Wenqi Shi, Yuanda Zhu, Tarun Naren, Monica Isgut, Ying Sha, Li Tong, Mitali Gupte, and May D. Wang, "Explainable artificial intelligence methods in combating pandemics: A systematic review.," *arXiv:2112.12705v2*, 2021. 4

[25] Jordan D. Fuhrman, Naveena Gorre, Qiyuan Hu, Hui Li, ssam El Naqa, and Maryellen L. Giger, "A review of explainable and interpretable ai with applications in covid-19 imaging.," *Medical Physics 49:1-14*, 2022. 4

[26] Leland McInnes, John Healy, and James Melville, "Umap: Uniform manifold approximation and projection for dimension reduction.," *arXiv:1802.03426v3*, 2020. 4

[27] Mohammadreza Zandehshahvar, Marly van Assen, Hossein Maleki, Yashar Kiarashi, Carlo N. De Cecco, and Ali Adibi, "Toward understanding covid-19 pneumonia: a deep-learning-based approach for severity analysis and monitoring the disease.," *arXiv:1802.03426v3*, 2020. 4

[28] Yujin Oh, Sangjoon Park, and Jong Chul Ye, "Deep learning covid-19 features on cxr using limited training data sets.," *arXiv:1802.03426v3*, 2020. 4

[29] Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M Zughaier, Muhammad Salman Khan, et al., "Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images," *Computers in biology and medicine*, vol. 132, pp. 104319, 2021. 4

[30] Muhammad EH Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, et al., "Can ai help in screening viral and covid-19 pneumonia?," *IEEE Access*, vol. 8, pp. 132665–132676, 2020. 4, 8

[31] Anas M Tahir, Muhammad EH Chowdhury, Amith Khandakar, Tawsifur Rahman, Yazan Qiblawey, Uzair Khurshid, Serkan Kiranyaz, Nabil Ibtehaz, M Sohel Rahman, Somaya Al-Maadeed, et al., "Covid-19 infection localization and severity grading from chest x-ray images," *Computers in biology and medicine*, vol. 139, pp. 105002, 2021. 4

[32] Linda Wang, Zhong Qiu Lin, and Alexander Wong, "Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images," *Scientific Reports*, vol. 10, no. 1, pp. 19549, Nov 2020. 4

[33] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan, "Supervised contrastive learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18661–18673, 2020. 6

[34] Farheen Ramzan, Muhammad Usman Ghani Khan, Asim Rehmat, Sajid Iqbal, Tanzila Saba, Amjad Rehman, and Zahid Mehmood, "A deep learning approach for automated diagnosis and multi-class classification of alzheimer's disease stages using resting-state fmri and residual neural networks," *Journal of medical systems*, vol. 44, no. 2, pp. 1–16, 2020. 7

[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. 6

[36] Yoshua Bengio, Aaron Courville, and Pascal Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013. 12

[37] Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie, "When does contrastive visual representation learning work?," *arXiv preprint arXiv:2105.05837*, 2021. 13

[38] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008. 14