Bone Suppression as a Transformation in Self-Supervised Learning for Tuberculosis (TB) Diagnosis from Chest X-Rays.

Luciano Meixieira and **Richard Klein**
School of Computer Science and Applied Mathematics
University of the Witwatersrand, Johannesburg, South Africa

Abstract—**Tuberculosis (TB) remains one of the world's deadliest diseases. Early detection is crucial but there are limitations and complications associated with current diagnostic methods. Therefore, there is a need for affordable, accessible alternatives. Despite the availability of unlabeled medical imaging data, the lack of publicly available labeled imaging data provides limitations. This paper posits that leveraging the unlabeled data through self-supervised learning (SSL) can be a potential solution. A prevalent question in applying SSL to medical imaging is the efficacy of standard data augmentations in capturing meaningful representations inherent to such data. Addressing this, our research contrasts the performance of domain-specific data augmentations against generic augmentations in SSL for TB classification from Chest X-Ray (CXR) images. Our primary focus is on bone suppression, a domain-specific technique that removes the skeletal structures from a CXR whilst leaving the rest of the image untouched. We examine the effects of the bone suppression transformation in 4 different SSL methods. Subsequent to the self-supervised pre-training, our models were evaluated in a linear and full model fine-tuning setting using a smaller labeled dataset. Our overall results, derived from various performance metrics, indicated an increase in performance when utilizing the bone suppression technique as a data augmentation. This finding not only underscores the potential of domain-specific augmentations in medical imaging but also paves the way for further exploration into novel augmentations tailored to specific diseases and domains [1].**

I. Introduction

Tuberculosis (TB) remains the world's deadliest individual pathogen, despite being both preventable and curable [1]. According to [1], the African and South-East Asian regions have experienced the most significant reported tuberculosis mortality rates, where South Africa has produced the 8th highest TB incidence rate globally. South Africa remains in the group of high TB-burdened countries and is one of 16 countries that account for 93% of the global TB burden [2].

Smear microscopy is the most prevalent diagnostic method for TB, however, it comes with limitations and diagnostic difficulties [5]. While mycobacterial culture offers improved sensitivity for detecting TB across various samples, the extended time required to obtain results causes delays in reporting outcomes [6, 7]. It is vital that TB is detected early as late detection leads to an increase in transmission risk, economic hardship, and risk of death [3]. TB can be detected through chest X-rays (CXRs). However, only a qualified radiologist will be able to make any diagnosis, and unfortunately, there are a low number of these professionals available, especially in high TB-burdened countries [4]. Therefore, the development of a classification model for the automated diagnosis of TB using CXRs is justified.

However, there are some drawbacks to creating this model. While acquiring large amounts of image data in the medical field is possible, labeling these images would involve the intervention of specialists, which would be costly and is sometimes not possible [8, 9]. Thus, there is a need to explore training methods using data in which there are no specific targets or labels specified. One particular method we will focus on is self-supervised learning (SSL). SSL is a subcategory of unsupervised learning and offers a method for training a model that utilizes the input data effectively to create its own supervision [10]. In other words, the model learns to extract patterns and make predictions from unlabeled data, often by employing data augmentations to create its own training signals.

In this paper, we investigate whether standard data augmentations in SSL, such as random cropping and color distortion, for regular images, lead to useful representations in the medical image space. We propose a new data augmentation technique for this task known as bone suppression and the performance of this SSL pretext task will be evaluated against the standard augmentations. Bone-suppressed images relate to removing the bony structures in CXRs whilst leaving the lung tissue and other features of the CXR intact [11]. The goal is to determine if pre-training a model using cutting-edge SSL techniques using bone suppression as a transformation, will outperform other standard data augmentations to demonstrate the importance of logical, domain-specific transformations in SSL when diagnosing TB from CXRs.

We have evaluated 4 diverse state-of-the-art (SOTA) SSL techniques using different sets of data augmentations. Please refer to Table I for a list of these techniques:

TABLE I: Self-Supervised Learning Techniques

SimCLR	BYOL	SwAV	DINO
[12]	[13]	[14]	[15]

[1] We release the code used that achieved these results here: https://github.com/LucianoMeixieira/TB-Augmentations-SSL.git

Our findings indicate that incorporating the bone suppression transformation during the self-supervised pre-training phase enhances the model's capability to classify TB from CXRs. This underscores the pivotal role of domain-specific knowledge in the selection of data augmentations for SSL, thereby influencing the efficacy of the generated supervisory signal within the model.

Our paper aims to add to the conversation regarding the significance of data augmentations in cutting-edge SSL pre-training methods for the task of TB classification using CXRs.

The remainder of this document will follow the following structure:

- In the Background and Related Work section, we showcase research from the relevant literature, essential for comprehending the context and specifics of this paper. We first examine residual networks, as this will be the main backbone architecture used for all our models. Additionally, we touch on transfer learning as this is an important concept to understand when fine-tuning our models for classification. Furthermore, We explore various SSL techniques and the process of bone suppression while relating them to the core task of automated diagnosis of TB using CXRs. Furthermore, we look at supervised approaches to this problem which include ensemble learning. Finally, we explain the justification for our research in terms of the past research that has been done in the field of SSL for medical image analysis.
- In the Methadology section, we portray the procedural steps taken to establish our conclusions and synthesize our findings. This section comprehensively covers the acquisition and curation of the datasets, the detailed implementation of various SSL methods for pre-training, and the protocols for fine-tuning. We conclude with a description of the evaluation metrics employed to measure the performance of our models.
- We showcase our findings and results in the Experiments and Results section. All experiments conducted in this section are described and include the effects that bone suppression has when utilized in a self-supervised domain.
- The Discussion section concisely summarizes the results obtained and provides an overall analysis of the effect that the bone suppression transform has.
- The Limitations and Future Work section discusses the limitations of our experiments which include memory and resource constraints which resulted in a sub-optimal parameter set used when pre-training using SSL. Furthermore, we discuss the potential for future work in this field, which includes research into creating other domain-specific augmentations for TB classification in CXRs as well as further experiments that could be conducted.
- Finally, in our Conclusion, we provide a summary of the paper.

II. Background and Related Work

This section will present our discoveries for works and relevant research on the topic. This section will present all the topics we intend to use for this paper and include Residual Networks, Transfer Learning used for fine-tuning pre-trained models, cutting-edge Self-Supervised Learning techniques, and finally, Bone Suppression data augmentation. Additionally, we discuss various papers and alternative methods for the computer-automated diagnosis of TB using CXRs. Finally, we provide justification for our research by delving into previous work that has been done in the realm of SSL for the medical imaging domain.

A. Residual Networks

A novel deep residual learning framework is presented in [16] that tackles an issue in deep Convolutional Neural Networks (CNNs), where an increase in the network depth leads to an increase in training and validation errors and subsequent performance decrease. The residual learning framework presented in [16] overcame these challenges by including skip connections between pairs of convolutional layers. The incorporation of these connections enables the network to learn residual functions in relation to the input of the layers, consequently facilitating optimization and mitigating the performance loss that occurs in deeper networks. The networks developed through this approach are referred to as Residual Networks (ResNets). The ResNet model can have a different number of layers. ResNet-n refers to a ResNet model that has n number of layers. This implementation has led to developments in several areas of image classification tasks [17]. Importantly, ResNet models have often been used to classify TB using CXRs, which can be reviewed in [18], as well as other pulmonary-related diseases such as Covid-19 [19]. Furthermore, ResNet models were used as the backbone architecture of all of the techniques we will be aiming to evaluate in Table I. Therefore, this would be a very suitable architecture for our task.

B. Transfer Learning

The concept of Transfer Learning and fine-tuning has previously been shown to be effective in the realm of medical imaging [22]. Transfer learning is a popular data-efficient approach that is often used when we do not have access to large amounts of training data. Transfer learning is a machine learning approach where a model that is pre-trained on a source task is adapted or fine-tuned for a target task [21]. Fine-tuning is a method that involves adjusting the model's weights, to adapt it to the target task.

The source and target tasks can either be related or unrelated, meaning they may or may not share similar structures. When the source and target tasks are related, it is anticipated that the pre-trained model has learned an important feature representation from the source data that is useful for the target task. In this case, fine-tuning the model on related data will cause minor adjustments in the model which is expected to increase performance when used as a classifier

[23]. When the source and target tasks are unrelated, the pre-trained model will have learned low-level and mid-level representations that are useful for adapting the model to various image classification tasks [24].

Grasping this concept is crucial because, in the following section, we will delve into the significance of Transfer Learning and fine-tuning pre-trained models with SSL techniques for applying the acquired representations to subsequent tasks, such as TB classification in CXRs.

C. Self-Supervised Learning

SSL provides a method to train a model utilizing the input data for its own supervised task and positively enhances all types of downstream tasks, including classification [10]. Meaning, SSL techniques do not require any labeled data to train a model effectively, which we have already established is a massive advantage in the realm of medical images. In essence, the SSL techniques we will evaluate in Table I follow similar high-level methodologies. A pretext task is defined which relies on patterns of the data to generate a supervised task for the model to learn. This task should be designed so that solving it requires learning meaningful features or representations of the data. This generally involves data augmentation to create various perspectives of the same instance, which act as the main learning representations. A model is then pre-trained on this pretext task using one of the techniques in Table I. When this training process is completed, the learned representations and features can be extracted and transferred to a downstream task, such as classification, using Transfer Learning. This process involves fine-tuning the model on a smaller dataset of labeled data specific to our target task in order to increase our model's performance.

This is beneficial to our research as we will be able to pre-train a model on a large dataset of unlabeled CXRs to learn the necessary data representations, and then fine-tune our model to specifically classify the presence of TB. However, [25] suggests that certain standard data augmentations pretext tasks may compromise the raw data contained in medical images. Therefore, it is crucial to develop appropriate pretext tasks and make use of intelligent data augmentation strategies based on specific features of medical image data [25]. This further justifies our research in looking at an alternative data augmentation method used in SSL, specifically targeted at CXRs.

In the remainder of this section, we will explore the various SOTA SSL techniques presented in Table I, and showcase their relevancy to our paper.

*1) **SimCLR**:* [12] proposes a simple and effective Self-Supervised framework for learning visual representations, known as SimCLR. This is a contrastive learning technique that utilizes data augmentations where 2 different augmented instances of the input image are generated and referred to as a positive pair. Negative pairs refer to augmented samples from different input images. A CNN is used as a base encoder, which extracts features from the positive pairs while a projection head maps these features into a latent space where a contrastive loss function can be applied. The authors use a Normalized Temperature-Scaled Cross-Entropy loss function, which encourages representations of positive pairs to be similar whilst penalizing similar representations of negative pairs.

Notably, [12] discovers the importance of data augmentations and explains that a composition of multiple transformations produces more effective representations and improves results. The authors found that a combination that included random cropping and color distortion produced the best results. However, we expect that this may not be the case when examining medical images, as we could lose valuable information. This further emphasizes how important appropriate data augmentations are in SSL.

*2) **BYOL**:* A subsequent paper by [13], argues that the contrastive approach presented by [12] was too sensitive to the choice of data augmentation and only performed well when color distortion was involved. [13] presented a different SSL technique for learning image representations without the need for negative pairs, referred to as Bootstrap Your Own Latent (BYOL). BYOL makes use of 2 neural networks known as 'online' and 'target' networks. The online network shares an identical structure with the target network, with the exception of an additional prediction layer. The online network is trained to predict the output of the target network by processing different augmented samples of the input image. The authors employ the mean squared error (MSE) loss function, which essentially demonstrates the discrepancy between the target network's description and the online network's approximation.

The authors argued that BYOL was more robust to various data augmentations as it is encouraged to maintain supplementary characteristics in its feature representation compared to SimCLR. However, the authors stressed that BYOL remains dependent on appropriate data augmentations that are specific to the domain application.

*3) **SwAV**:* [14] noted that the existing contrastive SSL methods relied on a large number of comparisons between features of positive pairs which is very computationally taxing. The authors propose an instance discrimination online clustering algorithm that takes advantage of the contrastive methods, without the computational inefficiencies. This method is known as SwAV. SwAV does not need to perform pairwise comparisons between all images, but rather, the model is trained to predict the cluster assignments of one instance of an input image using the assignments of another instance. This encourages the model to learn consistent and robust features across different augmentations of the same image.

Moreover, [14] notes the importance of data augmentations in SSL, and proposes a new transformation strategy known as "multi-crop". The authors note that comparing more views and augmentations during training improves the performance of the model, which is why the "multi-crop" transformation was developed. This method works by generating multiple cropped versions of the same image at different resolutions. This was done to expose the model to multiple different perspectives of the input image, improving the generalization of the model. This again emphasizes the importance of data augmentations to improve downstream task performance in SSL.

*4) **DINO**:* Knowledge distillation is the process of transferring the knowledge of a larger, more complex model, to a smaller, more efficient model that performs just as well as the larger model [26]. The authors of [15] proposed a self-distillation SSL method known as DINO. In practice, DINO is very similar to the work presented by [13], however, they use different loss functions. DINO was initially implemented to explore whether SSL can be adapted to Vision Transformers (ViTs) in order to compete with CNNs. However, the authors stress that DINO is flexible and can be implemented for both ViTs and CNNs seamlessly, and can specifically be applied to ResNets which the authors implemented for testing purposes. DINO involves training a student network to predict the output of a teacher network for different augmented views of the same image.

DINO was tested with various data augmentations using the multi-cropping transformation strategy that was implemented by [14]. The authors noted that utilizing this training method improved performance, along with a combination of data augmentations such as solarization, and color jittering. DINO was compared to the other SOTA SSL techniques and outperformed all of them, including the ones discussed in this section for the downstream task of classification. Therefore, it makes sense to test a new domain-specific data augmentation for medical images using the DINO technique.

D. Bone Suppression

Pinpointing subtle findings related to TB can be challenging for both radiologists and predictive models, especially when these findings appear in apical regions where lung parenchyma is hidden by ribs and clavicle [20]. This challenge may arise due to the two-dimensional nature of CXRs, which causes posterior and anterior bony structures to overlap with lung tissues. Moreover, the prominent edges of the bones may cover anomalies in the lung regions, making diagnosis more complicated. Therefore, [11] built various deep learning-based bone suppression models, that identified the occluding bone structures in CXRs whilst leaving the lung tissue and the rest of the CXR untouched. The best-performing model on several metrics was a Residual Network model which was called ResNet-BS (ResNet Bone Suppression) which successfully suppressed the bone structures while simultaneously keeping the output incredibly close to its ground truth. Please refer to Figure 1 for an example.

Notably, [11] discovered that a model trained on CXR data and fine-tuned on bone-suppressed images significantly outperformed a model fine-tuned on non-bone suppressed images. This concluded that the bone suppression technique helped the model to more efficiently identify patterns of X-ray data that were indicative of TB. Features of interest were more closely grouped together and therefore, it was easier to classify TB-related features.

This, therefore, justifies why it would be sensible to use this technique as a transformation in Self-Supervised Learning methods. The hope is that during the training process, the model will learn to ignore the bone structures in the data, leading to the model learning more valuable features, resulting in improved class relevance mapping localization, when it comes to classifying TB in CXR data.

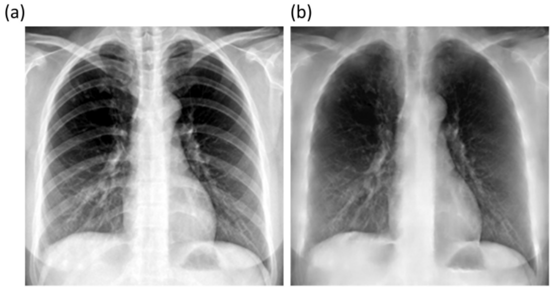

Fig. 1: Side-by-side comparison of a normal chest X-ray and its bone-suppressed counterpart produced by [11]. (a) represents a normal CXR and (b) represents the bone-suppressed counterpart.

E. Supervised Learning for TB Diagnosis

Another approach to our problem is Supervised learning, which describes training a model with fully labeled data as opposed to unlabeled data. Many Supervised approaches have been explored in an attempt to create a computer-aided diagnostic TB classification model. [18] comprehensively reviewed 54 peer-reviewed papers between the years 2016-2021 which attempted to screen TB using CXRs. This paper noted that deep learning techniques such as ensemble learning were preferred as deep features were automatically extracted from the input images as opposed to hand-crafting features that require expert analysis.

*1) **Ensemble Learning**:* Ensemble learning is a machine learning technique that amalgamates multiple models into one model, in order to improve the performance and generalization of predictions [27]. [28] proposed a TB detection model that utilizes ensemble learning to merge hand-crafted features with deep features extracted from a CNN. To extract the handcrafted features, the authors experimented with 2 different configurations of the Gabor filter, while the deep features were extracted using 7 pre-trained CNN models. The features extracted through each of these models were then used to train a classifier that represented the first phase of predictions. To achieve the second phase of predictions, the individual outputs were combined and used to train a new classifier through ensemble learning. Another paper by [29] ensembles two different CNN architectures and trains them on 2 different types of input images. The first type is a normal CXR image and the second is a canny edge detection CXR image. This is to generate variety in the features that are extracted from the different inputs. For each CNN architecture, two classifiers will be created from the two sets of images, meaning there will be four individual classifiers that can be ensembled. At the start of the classification phase, an individual CNN will be trained to detect TB. To achieve an ensembled model, a technique that averages the probability scores of the different models and combines them to produce a final output was utilized.

The performance of both ensemble learning methods outperformed the regular supervised models. Both ensembled models made use of different data augmentations in order to extract different features that a model could be trained on. This illustrates the positive effect data augmentations can have on the performance of a classifier which is how these methods relate to our project.

However, one criticism one could have with these models reviewed in [18] is that the datasets used to train and test the models are quite small. Therefore, we hypothesize that these models will not generalize well if pushed into production.

F. Self-Supervised Learning for Medical Images

Several recent studies have demonstrated the potential of SSL in the analysis of medical images [30, 31, 32]. Additionally, [25] offered a comprehensive review of SSL applications in the medical imaging domain, further underscoring its potential in this area. However, a notable research gap persists. No research has specifically probed into domain-specific data augmentation methods for CXRs or evaluated their impact on the performance of a Self-Supervised classifier tailored for TB detection. Although [11] evidenced the benefits of a domain-specific augmentation in supervised learning, its implications for SSL remain uncharted. Moreover, we hypothesize that common data augmentation techniques, like random cropping, which have been efficacious for natural images in SSL, might not be equally fruitful for nuanced medical images, such as CXRs, especially in the context of TB detection.

This brings us to a significant contribution by [33], which stands out as the most comprehensive of its kind on pathology image data. The study underlined the key distinctions between medical and natural images and argued that large-scale domain-aligned pre-training can potentially eclipse the outcomes of ImageNet pre-training [34]. To substantiate this, the researchers introduced a set of domain-specific techniques for the self-supervised pre-training phase. Their experiments revealed that these domain-specific augmentations significantly enhanced performance across various downstream tasks. The overarching conclusion was the paramount importance of domain knowledge in SSL pre-training; integrating this knowledge led to superior results compared to a naive application of SSL. This work provides a justification for our research and displays the potential that the bone suppression transformation can have in using SSL to classify TB in CXRs.

G. Conclusion

In this section, we have introduced literature that has investigated the topics of Residual Networks, Transfer Learning, Self-Supervised Learning, and Bone Suppression, which are fundamental to the understanding and proceeding of our paper. We have shown the relevance of these topics to our research and explained the background of these concepts. We have explored papers that are related to our research and showcased the importance of data augmentations and domain knowledge when analyzing medical images in a self-supervised setting. Finally, we verify that the goals of this paper are well-founded based on previous research and that our research is justified.

III. Methodology

This section provides a detailed explanation of the methods and approaches used in order to achieve our results. We explain the data collection and preparation steps, as well as the pre-training processes for each SSL method mentioned in Table I. We explain our fine-tuning process where we use both linear evaluation and full model-fine tuning and finally, discuss our evaluation and testing metrics.

A. Data Preparation and Collection

All the CXR images were converted to grayscale and the pixel data was converted to a floating-point representation in the range [0.0, 1.0]. Furthermore, the images were resized to 256×256 pixels in order to be passed into the bone suppression model[2]. Once an image has been preprocessed, it is passed through this model, after which both the original image and the corresponding bone-suppressed output are stored. It is important to note that we strictly worked with frontal CXRs. The following datasets were used:

1) The TBX11 dataset presented by [36] contains 11200 X-ray images categorized into 5 groups: Healthy, Sick but Non-TB, Active TB, Latent TB, and Uncertain TB. In addition, this dataset has included additional datasets that contain active-TB and healthy frontal CXRs [37, 38], taking the total to 12278 images and providing a diverse set of data. Due to this diversity, this dataset was used in the self-supervised pre-training phase for all the SSL methods. Therefore, the 5 categories of this dataset do not matter, as we will not be making use of any labels for this dataset.
2) [39] presented a dataset containing 3500 TB CXRs and 3500 normal CXRs. This dataset was used to fine-tune and evaluate our pre-trained models and therefore, labels for this dataset were created with 0 representing normal CXRs and 1 representing a CXR that contains TB. We randomly shuffled this data and used a 70/10/20 random split in order to create our training, validation, and testing sets.

B. Implementation Details for Self-Supervised Pre-Training

SSL demands a backbone that can effectively extract features from the different augmented views of the input images. The ResNet50 architecture serves as the backbone for all models, modified to accommodate grayscale images by utilizing a single-channel input. The output of this backbone is the output of the final average pooling layer, which has a feature dimension of 2048. To assess the impact of bone suppression techniques within SSL, we prepare 4 sets of data augmentations. Each SSL method will be coupled with all 4

[2] The code for this model was adapted from the following repository: https://github.com/sivaramakrishnan-rajaraman/CXR-bone-suppression.git

augmentation sets, resulting in a total of 16 distinct models. These data augmentation sets are delineated as follows:

1) **"Default"**: This set comprises the data augmentations used in the official implementation of the literature and will act as our **baseline**.
2) **"Bone Supp"**: This set contains exclusively bone-suppressed images, resized to 224×224 pixels, with bone suppression being the sole augmentation technique.
3) **"Bone Default"**: This set integrates the default augmentations applied to bone-suppressed images, thus incorporating bone suppression into the standard augmentation procedures.
4) **"New Combo"**: This set introduces a novel combination of data augmentations, which includes the application of bone suppression among others.

For more details regarding the above data augmentation strategies applied, refer to Appendix VII-A.

Owing to computational and memory constraints, it was impracticable to implement the SSL methods using the optimal parameters reported in the literature. Specifically the batch size and number of epochs. Nonetheless, informed by insights from the literature, we conducted a robust comparison of our models using a consistent but potentially sub-optimal parameter set. The loss curves for the respective self-supervised pre-training techniques can be found in Appendix VII-B. It is imperative to note that the primary objective of this research is not to achieve peak performance but to critically examine the influence of domain-specific data transformations on SSL efficacy. Hence, this comparative analysis prioritizes the assessment of the bone suppression technique within the SSL context over the pursuit of maximal accuracy. Detailed descriptions of the adaptations and implementations for each SSL method can be found below[3].

*1) **SimCLR**:* A base encoder backbone, with a 2-layer Multi-Layer Perceptron (MLP) projection head to project the representation into a 128-dimensional latent space. The non-linear MLP consisted of a linear layer followed by a rectified linear units (ReLU) [42] activation, with a final linear layer. We used the NT-Xent loss function described in [12] with a temperature of 0.5, optimized with the LARS[4] optimizer. We used square root learning rate scaling as [12] reveals that this is preferred when working with smaller batch sizes. Therefore, the learning rate used was 1.2 ($0.075 \times \sqrt{BatchSize}$) with a weight decay of 10^{-6}. Each model was trained with a batch size of 256 for 500 epochs. Moreover, we used a linear warmup for the first 10 epochs and decayed the learning rate with a cosine decay schedule without restarts. These parameters are used for all models pre-trained using the SimCLR technique. For an explanation of the data augmentation sets used, please refer to **SimCLR Data Augmentations**.

*2) **BYOL**:* A backbone encoder for both the online and target networks facilitates the extraction of feature representations. These representations are subsequently mapped to a lower-dimensional space using a MLP. The MLP configuration includes an initial linear layer with an output dimensionality of 4096, which is followed by batch normalization and ReLU activation. The process culminates with a final linear layer that reduces the output to a 256-dimensional vector. Due to BYOL's increased robustness, training with a smaller batch size should not significantly decrease the performance [13]. The authors provide guidelines for parameter adjustments under such conditions. Consequently, our experiments are conducted with a batch size of 256, adopting a base learning rate of 0.4. This is adjusted linearly in proportion to the batch size, using the formula $0.4 \times \frac{\text{BatchSize}}{256}$, alongside a weight decay parameter set at 1.5×10^{-6}. We employ the LARS optimizer coupled with a cosine decay learning rate schedule, without restarts, across 500 epochs. The schedule includes an initial warmup phase spanning 10 epochs. For the target network update mechanism, the exponential moving average coefficient, τ, is configured to 0.9995. To accommodate a reduction in batch size by a factor of N, we average the gradients over N sequential steps, synchronously updating the target network after every Nth step to maintain consistency in the training regimen. Finally, the mean squared error (MSE) loss function is used as described in the literature. These pre-training configurations are utilized for all data augmentation sets with further specifics found in **BYOL Data Augmentations**.

*3) **SwAV**:* Similarly to the works presented in [12, 13], a 2-layer MLP projection head is used on top of the features extracted from our backbone to project the output into a 128-feature dimension space. The MLP consists of an initial linear layer that projects the features into a 2048-D space, followed by batch normalization and a ReLU activation. Thereafter, this representation is projected into the 128-D space. This is trained with a batch size of 256 for 200 epochs with a queue length of 3840 ($15 \times BatchSize$) to simulate large batches [14]. This queue is composed of feature representations from previous batches and is only used after 15 epochs of training. We use a weight decay of 10^{-6} and a LARS optimizer with a learning rate of 0.6. A cosine learning date decay schedule is used to achieve a final value of 0.0006. The temperature parameter, τ, is set to 0.1 with the SinkHorn regularization parameter, ϵ, set to 0.05. To help the initial optimization, the 3000 prototypes are frozen during the first epoch of training. The loss function used during this process is the same loss function designed in the literature. For the multi-crop data augmentation strategy, we produce 2 global views with 6 local views with the specific data augmentation sets described in **SwAV Data Augmentations**.

*4) **DINO**:* The student and teacher networks are composed of a backbone network and a projection head. The projection head consists of a 3-layer MLP. The first and second layers comprise hidden dimensions of 2048 with Gaussian error linear units (GELU) activations. The third layer reduces

[3] The adaptation and detailed implementation of each SSL method was guided by the resources available at the following repository: https://github.com/giakou4/pyssl

[4] We utilize the Layer-wise Adaptive Rate Scaling (LARS) optimizer implementation from the following repository: https://github.com/kakaobrain/torchlars/tree/master

the dimensionality to a 256-dimensional latent space without any non-linear activation. Following the transformations,ℓ_2 normalization is applied along the last dimension. The final output is obtained by passing the normalized vector through a weight-normalized linear layer, initialized such that the norm of the weights is set to 1 and frozen to preclude gradient updates. Weights for the linear layers are initialized using a truncated normal distribution with a standard deviation of 0.02, while biases are initialized to zero, ensuring a standardized starting point for learning [15]. The models are pre-trained for 100 epochs with a batch size of 128 with the AdamW optimizer [43]. The learning rate is linearly ramped up during the first 10 epochs to its base value of 0.00025 determined by: $(0.0005 \times \frac{\text{BatchSize}}{256})$. After this warmup, the learning rate is decayed with a cosine schedule. The weight decay also follows a cosine schedule from 0.04 to 0.4. The temperature τ_s is set to 0.1 and τ_t is set to 0.04. The smoothing parameter for the teacher, m, is set to 0.9 with λ set to 0.996. Similar to SwAV, we utilize a multi-crop strategy for our data augmentation sets with 2 global views and 6 local views. More information about the data augmentations used can be found in **DINO Data Augmentations**.

C. *Implementation Details for Fine-Tuning*

This process entails loading in the backbone of the pre-trained model and adding a logistic regression classifier to this backbone. The weights of the classification layer are sampled from a normal distribution with mean = 0 and standard deviation = 0.01. When fine-tuning the model, the input images are randomly cropped and resized to 224×224 with a scale = (0.08, 1.0). Thereafter, they are flipped horizontally with a probability of 0.5. Our models are fine-tuned in 2 different settings. These processes are described as follows:

1) **Linear Evaluation** refers to freezing the weights of our model before retraining the last logistic classification layer on our strictly TB-positive and TB-negative data [40, 41]. This procedure is performed to assess the effectiveness of the features learned by the model.
2) **Full Model Fine-Tuning** refers to fine-tuning the entire model on the strictly TB-positive and TB-negative data. Therefore, the weights of the model are fine-tuned slightly in order for the model to specifically classify whether a CXR is infected with TB.

A hyperparameter sweep was conducted by taking the center crop during the evaluation and minimizing the validation loss. The fine-tuning configurations are displayed in Table II.

D. *Evaluation Metrics*

The fine-tuned models from Implementation Details for Fine-Tuning are loaded in and evaluated. The following metrics are used to analyze our models: Accuracy, Precision, Recall (Sensitivity), Specificity, and Area Under the Curve (AUC). Additionally, the confusion matrices and AUC-ROC curves have been constructed and can be analyzed in Appendix VII-C.

IV. Experiments and Results

In this section, we analyze the performance of each SSL method under the influence of our distinct data augmentation strategies, evaluating the effect of the bone suppression transformation. For every SSL approach, we conduct an intra-method comparison to determine the most effective augmentation set. Each augmentation strategy is systematically evaluated according to the two fine-tuning paradigms described in Implementation Details for Fine-Tuning. This analysis aims to pinpoint the augmentation strategy that yields the best results for each SSL method and to decipher the impact of the bone suppression transformation. The results achieved from **SimCLR**, **BYOL**, **SwAV**, and **DINO** can be seen in the respective tables Table III, Table IV, Table V, and Table VI. Note that the **green** text indicates the best value obtained for that specific metric. Furthermore, the relevant confusion matrices and Receiver Operating Characteristic (ROC) curves can be found in Appendix VII-C.

A. *SimCLR Results and Analysis*

TABLE III: SimCLR Results

Augmentation Set	Accuracy	Precision	Recall	Specificity	AUC
Linear Evaluation:					
Default	83	86	80	87	83
Bone Supp	80	80	82	79	80
Bone Default	**85**	**89**	80	**90**	**89**
New Combo	84	79	**92**	75	84
Fine-Tuning:					
Default	84	86	81	87	84
Bone Supp	75	69	**92**	58	75
Bone Default	**86**	**88**	83	**89**	**86**
New Combo	85	85	85	84	85

In the SimCLR results, the ”Bone Default” augmentation strategy stands out with the highest accuracy, precision, specificity, and AUC in both linear evaluation and full fine-tuning settings. This indicates a robust capability in identifying TB in CXRs with a low rate of false positives and high confidence in the model’s predictive power. In contrast, the ”Default” set, which lacks the bone suppression transformation, consistently underperforms relative to ”Bone Default”, underscoring the effectiveness of bone suppression in improving overall performance. Notably, the ”New Combo” set exhibited the highest recall during linear evaluation, and ”Bone Supp” matched this during full fine-tuning, suggesting their effectiveness in capturing true TB cases despite an increase in false positives. However, these 2 sets did not surpass the balanced performance of the ”Bone Default” set. Fine-tuning generally enhanced model performance slightly, suggesting that while pre-trained weights provided a solid foundation, additional task-specific training was advantageous. The consistency of the ”Bone Default” set’s performance suggests that the features learned with this augmentation set are particularly well-suited for TB detection, making it the most reliable choice amongst the evaluated sets for SimCLR. These results collectively affirm the value of incorporating bone suppression in the

TABLE II: Fine-Tuning Configurations

	Backbone	Learning Rate	Weight Decay	Epoch	Batch Size	Momentum	Optimizer	Scheduler
Linear Evaluation:								
SimCLR	Encoder	0.005	N/A	140	256	0.9	Nesterov SGD	N/A
BYOL	Online Encoder	0.005	N/A	110	256	0.9	Nesterov SGD	N/A
SwAV	Encoder	0.01	7.367×10^{-5}	150	256	0.9	SGD	Cosine Decay
DINO	Teacher	0.00064	N/A	300	64	0.9	SGD	Cosine Decay
Fine-Tuned:								
SimCLR	Encoder	0.000625	N/A	110	256	0.9	Nesterov SGD	N/A
BYOL	Online Encoder	0.005	N/A	90	256	0.9	Nesterov SGD	N/A
SwAV	Encoder	0.01	7.584×10^{-5}	350	128	0.9	SGD	Cosine Decay
DINO	Teacher	0.00036	N/A	250	64	0.9	SGD	Cosine Decay

augmentation pipeline, which is instrumental in refining the model's ability to distinguish TB features in the presence of complex bone structures.

B. BYOL Results and Analysis

TABLE IV: BYOL Results

Augmentation Set	Accuracy	Precision	Recall	Specificity	AUC
Linear Evaluation:					
Default	72	67	**88**	55	71
Bone Supp	**79**	**78**	81	**77**	**79**
Bone Default	58	56	72	44	58
New Combo	72	75	67	**77**	72
Fine-Tuning:					
Default	**68**	**93**	39	**97**	**68**
Bone Supp	66	**93**	35	**97**	66
Bone Default	58	55	**92**	23	58
New Combo	62	85	29	95	62

The BYOL results highlight the effectiveness of bone suppression in detecting TB in CXRs. The ”Bone Supp” set demonstrated superior accuracy, precision, specificity, and AUC during the linear evaluation phase, suggesting its robustness for balanced TB detection. Meanwhile, the ”Default” set achieved the highest recall, indicating its strength in identifying the majority of true TB cases but at the expense of a significant number of false positives, as evidenced by its low specificity. Notably, during full model fine-tuning, ”Default” emerged with the best accuracy and AUC, as well as the highest precision and specificity, suggesting a refined ability to correctly identify TB cases. However, this came with a poor recall, suggesting a high number of false negatives. ”Bone Supp” mirrored these precision and specificity metrics, although with slightly lower accuracy, recall, and AUC. The ”Bone Default” set revealed a high recall with the trade-off of lowest specificity, indicating a risk of increased false positives. This contrast between linear evaluation and full fine-tuning underscores the impact of augmentation strategies on model performance and the importance of selecting the appropriate strategy based on the desired diagnostic outcome. These findings indicate that while ”Bone Supp” is generally robust in linear evaluation, full model fine-tuning can shift performance dynamics, emphasizing the necessity of strategic augmentation selection to balance recall with precision and specificity for practical TB screening. This balance is crucial as it reflects the nuanced role of augmentation strategies in model performance, which extends beyond the baseline provided by pre-trained weights. Overall, the ”Bone Supp” augmentation set demonstrated the most consistent performance across various metrics during the linear evaluation, indicating that the features it highlights are apt for initial TB detection. This consistency underlines the significance of bone suppression in enhancing detection reliability.

C. SwAV Results and Analysis

TABLE V: SwAV Results

Augmentation Set	Accuracy	Precision	Recall	Specificity	AUC
Linear Evaluation:					
Default	79	85	**71**	87	79
Bone Supp	71	85	51	91	71
Bone Default	**80**	**90**	68	**92**	**80**
New Combo	79	88	68	90	79
Fine-Tuning:					
Default	**80**	87	**71**	89	**80**
Bone Supp	73	86	56	**91**	73
Bone Default	**80**	**89**	70	**91**	**80**
New Combo	**80**	88	70	**91**	**80**

During linear evaluation for SwAV, the ”Bone Default” set achieved the highest accuracy, precision, specificity, and AUC. Notably, it outperformed the ”Default” set, which lacks bone suppression, on all metrics except recall. This superior performance of ”Bone Default” over ”Default” underscores the potential value of bone suppression combined with other transformations. However, the ”Bone Supp” set, which focuses solely on bone suppression, lagged behind in recall and accuracy, suggesting that while bone suppression aids in specificity, it might not be as effective on its own without additional augmentations. In full fine-tuning, the results are more homogenized, with ”Default,” ”Bone Default,” and ”New Combo” all reaching an accuracy and AUC of 80%. The ”Default” set maintains the highest recall, as in the linear evaluation, indicating its consistent ability to detect the majority of true

TB cases without the aid of bone suppression. This consistency in recall, coupled with competitive precision and specificity scores, suggests that while bone suppression is beneficial, the ”Default” set remains a strong baseline. The ”Bone Supp” set, even after fine-tuning, does not surpass ”Default” in recall or accuracy but maintains a high specificity of 91%. The ”Bone Default” and ”New Combo” sets, both incorporating bone suppression, share top scores in precision and specificity along with the ”Default” set, reinforcing that bone suppression, when used alongside other augmentations, improves the model’s ability to correctly identify negative cases and true positives. Overall, these SwAV results illustrate that bone suppression, while effective, is not the sole determinant of a model’s success in TB detection. Its impact seems most pronounced when combined with other transformations, as shown by the performance of the ”Bone Default” and ”New Combo” sets. This analysis highlights the nuanced role of bone suppression—it may not significantly boost performance on its own but can enhance the overall efficacy when integrated with a comprehensive set of augmentations.

D. DINO Results and Analysis

TABLE VI: DINO Results

Augmentation Set	Accuracy	Precision	Recall	Specificity	AUC
Linear Evaluation:					
Default	50	51	54	46	50
Bone Supp	66	85	41	**93**	67
Bone Default	80	79	**83**	77	80
New Combo	**87**	**91**	**83**	92	**87**
Fine-Tuning:					
Default	44	45	57	30	43
Bone Supp	66	**89**	38	**95**	67
Bone Default	76	70	**92**	61	76
New Combo	**85**	**89**	81	90	**85**

The DINO framework’s results compellingly illustrate the impact of bone suppression in TB detection from CXRs. Notably, all augmentation sets incorporating bone suppression outperform the ”Default” set, which lacks this transformation, across almost all metrics. During linear evaluation, ”New Combo,” which includes bone suppression, achieves the best overall performance with the highest accuracy, precision, specificity, and AUC. This underscores the transformation’s utility in enhancing model discernment capabilities. The ”Bone Default” set follows, with strong results, particularly in recall, demonstrating its efficiency in identifying true positive cases. For full model fine-tuning, the augmentation sets with bone suppression again display superior outcomes, with ”New Combo” leading and ”Bone Supp” showing exceptional specificity. This indicates that bone suppression is particularly effective in reducing false positives, a critical aspect of diagnostic imaging. The ”Default” set’s substantially lower performance in both fine-tuning settings, despite having the highest recall in fine-tuning, suggests it may identify more true TB cases but is prone to a higher rate of false positives, evidenced by its low specificity and overall lower metrics. Overall, bone suppression’s inclusion consistently correlates with improved model performance. These results serve as a strong endorsement for the role of bone suppression in augmenting the detection quality of TB from CXRs, with the caveat that its absence does not necessarily prevent a model from identifying TB cases but may impair overall diagnostic accuracy.

V. Discussion

Our comprehensive analysis across different SSL frameworks—SimCLR, BYOL, SwAV, and DINO—reveals intriguing insights into the role of domain-specific augmentations, particularly bone suppression, in diagnosing TB from CXRs. Throughout the experiments, augmentation strategies that included bone suppression consistently outperformed the ”Default” sets, which did not employ this technique, across a range of metrics.

In SimCLR, the ”Bone Default” augmentation emerged as a superior strategy, achieving balanced performance in detecting TB. For BYOL, the incorporation of bone suppression as an exclusive domain-specific augmentation yielded promising results during linear evaluation, suggesting initial benefits. However, its relative underperformance compared to ”Default” for full model fine-tuning hints that bone suppression alone might not capture the full spectrum of information necessary for optimal TB detection. This observation suggests that integrating additional domain-specific augmentations could potentially unlock further benefits, emphasizing the importance of a more diversified augmentation strategy rather than relying on a single domain-specific transformation.

SwAV’s evaluation demonstrated that bone suppression is most effective when combined with other augmentations. While the ”Bone Supp” set showed high specificity, it was the comprehensive augmentations in ”Bone Default” and ”New Combo” that led to top performance, suggesting that bone suppression alone is not a panacea but rather a part of a holistic augmentation approach.

The DINO framework’s results unequivocally confirm the vital role of bone suppression in TB detection from CXRs. Augmentation sets with bone suppression significantly outperformed the ”Default” set, indicating that bone suppression is essential for improved performance. The ”New Combo” set, inclusive of bone suppression, delivered the best performance across all metrics, underlining the transformative effect of this technique in enhancing the model. The consistency in the superiority of bone suppression-inclusive sets across the board highlights its importance as a core augmentation for advancing TB detection in SSL models.

Importantly, no single SSL method emerged as universally superior for TB classification, indicating that the quest for optimal performance is not solely about the choice of SSL algorithm but is intricately tied to the augmentation strategies employed. The inclusion of bone suppression, a domain-specific augmentation, consistently led to superior performance, solidifying its role in developing robust models for medical imaging tasks.

In conclusion, while bone suppression is a critical factor in enhancing TB detection from CXRs, it is most effective when utilized within a comprehensive set of augmentations. These findings advocate for a nuanced approach to SSL pre-training where the combination of domain-specific augmentations is tailored to the specific needs of the task at hand, rather than seeking a one-size-fits-all solution within the myriad of SSL techniques.

VI. Limitations and Future Work

While the findings from this paper advance our understanding and application of domain-specific augmentations, there are inherent limitations that must be acknowledged. Below we outline specific constraints encountered and suggest avenues for future research to surmount these limitations, thus advancing SSL in medical imaging diagnostics.

- The pre-training of our SSL models did not employ the larger batch sizes and extended number of epochs that are often recommended in the literature. The benefits of scaling up these parameters could yield more optimal results than those presented.
- Our choice of parameters for SSL pre-training, although informed by their respective papers, may not be ideally suited for our specific task. Future studies should include a thorough hyperparameter optimization phase.
- The results reported reflect single pre-training and fine-tuning cycles; averaging over multiple runs could provide more robust and reliable findings.
- While DINO supports implementation with ViTs, our experiments were confined to a ResNet architecture. Exploring ViTs could reveal different aspects of model behavior.
- A comparative analysis between different SSL methods was outside the scope of this study, presenting a fertile ground for subsequent research.
- The TB datasets used were relatively small in size; accessing and employing larger datasets could potentially boost model performance.
- No comparative experiments were performed to evaluate the SSL models against supervised counterparts initialized with random or ImageNet[5] weights.

The aforementioned limitations not only highlight the scope of our current study but also chart a course for future research endeavors. There exists a considerable opportunity for the development of new domain-specific data augmentations, specially designed for TB detection in CXRs. Our findings suggest that relying solely on bone suppression to provide domain knowledge does result in a performance increase. However, they also indicate that bone suppression needs other data augmentations in order to further improve performance. Hence, the creation of a comprehensive set of domain-specific augmentations, inclusive of but not limited to bone suppression, represents a pivotal next step for improving TB diagnosis through SSL.

VII. Conclusion

Tuberculosis (TB), predominantly a lung infection, persists as a global health challenge, underlining the imperative for efficient diagnostic mechanisms. The development of an automated, computer-aided diagnostic system for TB is not just a technological advancement but a necessity for accessible healthcare. In the realm of machine learning, the scarcity of labeled medical imaging data contrasts with the abundance of unlabeled data. Self-supervised learning (SSL) leverages this unlabeled data, enabling models to self-generate training signals through strategic data augmentations.

This study ventured into relatively uncharted territory by evaluating the efficacy of domain-specific data augmentations in SSL, specifically through the implementation of bone suppression in chest X-rays (CXR). Bone suppression, by isolating critical pulmonary features, enhances the model's interpretive accuracy, thereby outperforming standard augmentation techniques in our experiments. Our findings advocate for bone suppression's integration as a transformational tool within SSL frameworks rather than as a direct classification modality.

Conclusively, our research underscores the importance of domain-specificity in SSL for medical imaging. The integration of nuanced augmentations tailored to the complexities of medical data promises substantial strides in automated disease detection. In practice, this implies more accurate models that could support clinicians in diagnosing TB more rapidly and reliably, potentially saving lives by enabling earlier intervention. Looking forward, the adoption and refinement of such domain-specific SSL transformations could profoundly transform medical diagnostics, elevating the standard of care in communities worldwide.

Acknowledgments

The author would like to extend his sincere gratitude to his supervisor, Professor Richard Klein, whose guidance and support have been invaluable throughout the course of this research. Prof. Klein's expertise in the field has been instrumental in shaping both the direction and success of this paper. Moreover, the author wishes to express appreciation to the staff of the Wits Cluster. The computational aspects of this work benefited significantly from the High-Performance Computing infrastructure availed by the Mathematical Sciences Support unit at the University of the Witwatersrand. Their technical expertise and support have been invaluable in facilitating the extensive computational tasks required for this research.

Declaration

I, Luciano Aguiar Meixieira, hereby declare the contents of this paper to be my own work. This paper is submitted for the degree of Bachelor of Science with Honours in Computer Science at the University of the Witwatersrand. This work has not been submitted to any other university, or for any other degree.

[5] The ImageNet data set is presented by [35]

Appendix

A. Data Augmentation Details

This section describes the 4 data augmentation strategies used for each SSL method.

*1) **SimCLR Data Augmentations**:* Below we showcase the specific details for each data augmentation used in each augmentation set for SimCLR in Table VII and Table VIII. Please note that I and I' represent two different augmented views.

TABLE VII: SimCLR Default and Bone Supp Data Augmentation Sets

Default			Bone Supp		
Parameter	I	I'	Parameter	I	I'
Random Resized Crop Probability	1.0	1.0	Resized Size	224	224
Random Resized Crop Size	224	224	Bone Suppression Probability	0.0	1.0
Random Resized Crop Scale	(0.08, 1.0)	(0.08, 1.0)			
Horizontal Flip Probability	0.5	0.5			
Colour Jitter Probability	0.8	0.8			
Colour Jitter Brightness	0.8	0.8			
Colour Jitter Contrast	0.8	0.8			
Gaussian Blur Probability	0.5	0.5			
Gaussian Blur Kernel Size	23	23			
Gaussian Blur Sigma	(0.1, 2.0)	(0.1, 2.0)			

TABLE VIII: SimCLR Bone Default and New Combo Data Augmentation Sets

Bone Default			New Combo		
Parameter	I	I'	Parameter	I	I'
Random Resized Crop Probability	1.0	1.0	Resized Size	224	224
Random Resized Crop Size	224	224	Bone Suppression Probability	1.0	1.0
Random Resized Crop Scale	(0.08, 1.0)	(0.08, 1.0)	Horizontal Flip Probability	0.5	0.5
Horizontal Flip Probability	0.5	0.5	Gaussian Blur Probability	0.5	0.5
Colour Jitter Probability	0.8	0.8	Gaussian Blur Kernel Size	23	23
Colour Jitter Brightness	0.8	0.8	Gaussian Blur Sigma	(0.1, 2.0)	(0.1, 2.0)
Colour Jitter Contrast	0.8	0.8			
Gaussian Blur Probability	0.5	0.5			
Gaussian Blur Kernel Size	23	23			
Gaussian Blur Sigma	(0.1, 2.0)	(0.1, 2.0)			
Bone Suppression Probability	1.0	1.0			

*2) **BYOL Data Augmentations**:* Below we showcase the specific details for each data augmentation used in each augmentation set for BYOL in Table IX and Table X. Please note that I and I' represent two different augmented views.

TABLE IX: BYOL Default and Bone Supp Data Augmentation Sets

Default			Bone Supp		
Parameter	I	I'	Parameter	I	I'
Random Resized Crop Probability	1.0	1.0	Resized Size	224	224
Random Resized Crop Size	224	224	Bone Suppression Probability	0.0	1.0
Random Resized Crop Scale	(0.08, 1.0)	(0.08, 1.0)			
Random Resized Crop Ratio	(3.0/4.0,4.0/3.0)	(3.0/4.0,4.0/3.0)			
Random Resized Crop Interpolation	Bicubic	Bicubic			
Horizontal Flip Probability	0.5	0.5			
Colour Jitter Probability	0.8	0.8			
Colour Jitter Brightness	0.4	0.4			
Colour Jitter Contrast	0.4	0.4			
Gaussian Blur Probability	1.0	0.1			
Gaussian Blur Kernel Size	23	23			
Gaussian Blur Sigma	(0.1, 2.0)	(0.1, 2.0)			
Solarization Probability	0.0	0.2			
Solarization Threshold	N/A	0.5			

TABLE X: BYOL Bone Default and New Combo Data Augmentation Sets

Bone Default			New Combo		
Parameter	I	I'	Parameter	I	I'
Random Resized Crop Probability	1.0	1.0	Resized Size	224	224
Random Resized Crop Size	224	224	Bone Suppression Probability	1.0	1.0
Random Resized Crop Scale	(0.08, 1.0)	(0.08, 1.0)	Horizontal Flip Probability	0.5	0.5
Random Resized Crop Ratio	(3.0/4.0,4.0/3.0)	(3.0/4.0,4.0/3.0)	Gaussian Blur Probability	1.0	0.1
Random Resized Crop Interpolation	Bicubic	Bicubic	Gaussian Blur Kernel Size	23	23
Horizontal Flip Probability	0.5	0.5	Gaussian Blur Sigma	(0.1, 2.0)	(0.1, 2.0)
Colour Jitter Probability	0.8	0.8	Solarization Probability	0.0	0.2
Colour Jitter Brightness	0.4	0.4	Solarization Threshold	N/A	0.5
Colour Jitter Contrast	0.4	0.4			
Gaussian Blur Probability	1.0	0.1			
Gaussian Blur Kernel Size	23	23			
Gaussian Blur Sigma	(0.1, 2.0)	(0.1, 2.0)			
Solarization Probability	0.0	0.2			
Solarization Threshold	N/A	0.5			
Bone Suppression Probability	1.0	1.0			

*3) **SwAV Data Augmentations**:* Below we showcase the specific details for each data augmentation used in each augmentation set for SwAV in Table XI and Table XII. Please note that G represents one global image, G' represents another global image, and L represents one of the local images used in the multi-crop strategy.

TABLE XI: SwAV Default and Bone Supp Data Augmentation Sets

Default				Bone Supp			
Parameter	G	G'	L	Parameter	G	G'	L
Random Resized Crop Probability	1.0	1.0	1.0	Resized Size	224	224	N/A
Random Resized Crop Size	224	224	96	Random Resized Crop Probability	0.0	0.0	1.0
Random Resized Crop Scale	(0.14, 1.0)	(0.14, 1.0)	(0.05, 0.14)	Random Resized Crop Size	N/A	N/A	96
Color Jitter Probability	0.8	0.8	0.8	Random Resized Crop Scale	N/A	N/A	(0.05, 0.14)
Color Jitter Brightness	0.8	0.8	0.8	Bone Suppression Probability	1.0	0.0	1.0
Color Jitter Contrast	0.8	0.8	0.8				
Horizontal Flip Probability	0.5	0.5	0.5				
Gaussian Blur Probability	0.5	0.5	0.5				
Gaussian Blur Kernel Size	23	23	10.6				
Gaussian Blur Sigma	(0.1, 2.0)	(0.1, 2.0)	(0.1, 2.0)				

TABLE XII: SwAV Bone Default and New Combo Data Augmentation Sets

Bone Default				New Combo			
Parameter	G	G'	L	Parameter	G	G'	L
Random Resized Crop Probability	1.0	1.0	1.0	Resized Size	224	224	N/A
Random Resized Crop Size	224	224	96	Random Resized Crop Probability	0.0	0.0	1.0
Random Resized Crop Scale	(0.14, 1.0)	(0.14, 1.0)	(0.05, 0.14)	Random Resized Crop Size	N/A	N/A	96
Color Jitter Probability	0.8	0.8	0.8	Random Resized Crop Scale	N/A	N/A	(0.05, 0.14)
Color Jitter Brightness	0.8	0.8	0.8	Gaussian Blur Probability	0.5	0.5	0.5
Color Jitter Contrast	0.8	0.8	0.8	Gaussian Blur Kernel Size	23	23	10.6
Horizontal Flip Probability	0.5	0.5	0.5	Gaussian Blur Sigma	(0.1, 2.0)	(0.1, 2.0)	(0.1, 2.0)
Gaussian Blur Probability	0.5	0.5	0.5	Horizontal Flip Probability	0.5	0.5	0.5
Gaussian Blur Kernel Size	23	23	10.6	Bone Suppression Probability	1.0	1.0	1.0
Gaussian Blur Sigma	(0.1, 2.0)	(0.1, 2.0)	(0.1, 2.0)				
Bone Suppression Probability	1.0	1.0	1.0				

*4) **DINO Data Augmentations**:* Below we showcase the specific details for each data augmentation used in each augmentation set for DINO in Table XIII and Table XIV. Please note that G represents one global image, G' represents another global image, and L represents one of the local images used in the multi-crop strategy.

TABLE XIII: DINO Default and Bone Supp Data Augmentation Sets

Default				Bone Supp			
Parameter	G	G'	L	Parameter	G	G'	L
Random Resized Crop Probability	1.0	1.0	1.0	Resize Size	224	224	N/A
Random Resized Crop Size	224	224	96	Random Resized Crop Probability	0.0	0.0	1.0
Random Resized Crop Scale	(0.08, 1.0)	(0.08, 1.0)	(0.05, 1.0)	Random Resized Crop Size	N/A	N/A	96
Random Resized Crop Interpolation	Bicubic	Bicubic	Bicubic	Random Resized Crop Scale	N/A	N/A	(0.05, 1.0)
Horizontal Flip Probability	0.5	0.5	0.5	Random Resized Crop Interpolation	N/A	N/A	Bicubic
Color Jitter Probability	0.8	0.8	0.8	Bone Suppression Probability	1.0	0.0	1.0
Color Jitter Brightness	0.4	0.4	0.4				
Color Jitter Contrast	0.4	0.4	0.4				
Gaussian Blur Probability	1.0	0.1	0.5				
Gaussian Blur Kernel Size	23	23	23				
Gaussian Blur Sigma	(0.1, 2.0)	(0.1, 2.0)	(0.1, 2.0)				
Solarization Probability	0.0	0.2	0.0				
Solarization Threshold	N/A	0.5	N/A				

TABLE XIV: DINO Bone Default and New Combo Data Augmentation Sets

Bone Default				New Combo			
Parameter	G	G'	L	Parameter	G	G'	L
Random Resized Crop Probability	1.0	1.0	1.0	Resize Size	224	224	N/A
Random Resized Crop Size	224	224	96	Random Resized Crop Probability	0.0	0.0	1.0
Random Resized Crop Scale	(0.08, 1.0)	(0.08, 1.0)	(0.05, 1.0)	Random Resized Crop Size	N/A	N/A	96
Random Resized Crop Interpolation	Bicubic	Bicubic	Bicubic	Random Resized Crop Scale	N/A	N/A	(0.05, 1.0)
Horizontal Flip Probability	0.5	0.5	0.5	Random Resized Crop Interpolation	N/A	N/A	Bicubic
Color Jitter Probability	0.8	0.8	0.8	Bone Suppression Probability	1.0	0.0	1.0
Color Jitter Brightness	0.4	0.4	0.4	Horizontal Flip Probability	0.5	0.5	0.5
Color Jitter Contrast	0.4	0.4	0.4	Gaussian Blur Probability	1.0	0.1	0.5
Gaussian Blur Probability	1.0	0.1	0.5	Gaussian Blur Kernel Size	23	23	23
Gaussian Blur Kernel Size	23	23	23	Gaussian Blur Sigma	(0.1, 2.0)	(0.1, 2.0)	(0.1, 2.0)
Gaussian Blur Sigma	(0.1, 2.0)	(0.1, 2.0)	(0.1, 2.0)	Solarization Probability	0.0	0.2	0.0
Solarization Probability	0.0	0.2	0.0	Solarization Threshold	N/A	0.5	N/A
Solarization Threshold	N/A	0.5	N/A				
Bone Suppression Probability	1.0	0.0	1.0				

B. Pre-Training Loss Curves

Below we showcase the pre-training losses for each augmentation strategy in SimCLR, BYOL, SwAV, and DINO in Figure 2, Figure 3, Figure 4, and Figure 5.

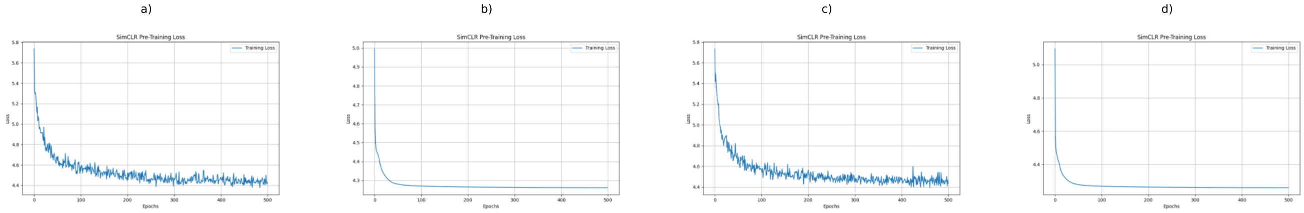

Fig. 2: SimCLR Pre-Training Loss Curve. (a) Default, (b) Bone Supp, (c) Bone Default, (d) New Combo.

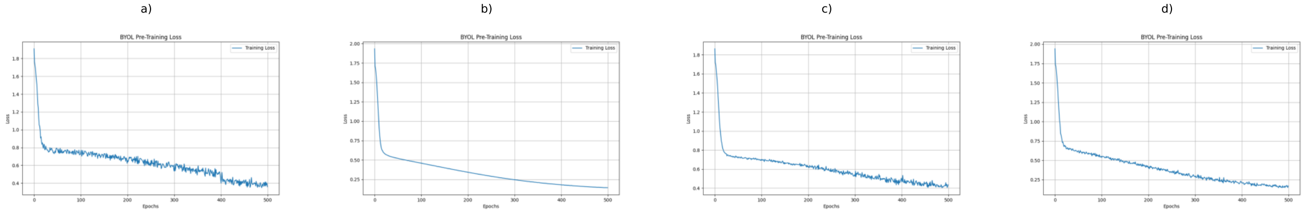

Fig. 3: BYOL Pre-Training Loss Curve. (a) Default, (b) Bone Supp, (c) Bone Default, (d) New Combo.

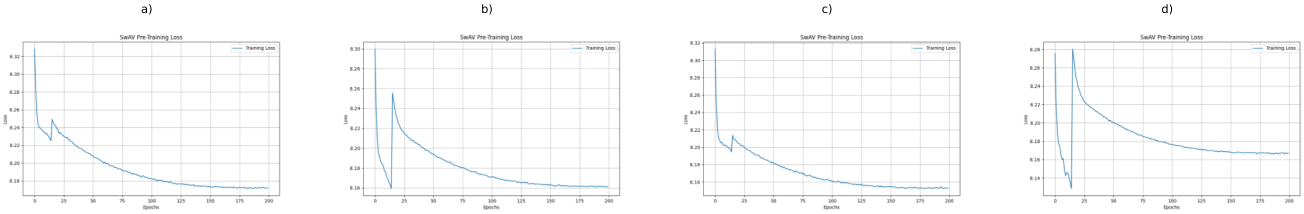

Fig. 4: SwAV Pre-Training Loss Curve. (a) Default, (b) Bone Supp, (c) Bone Default, (d) New Combo.

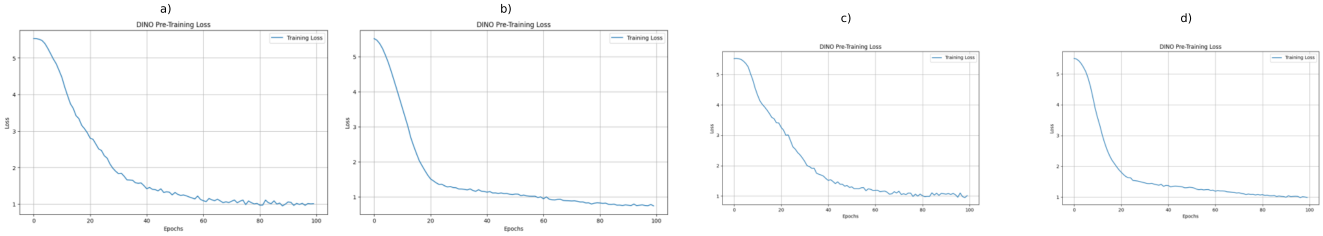

Fig. 5: DINO Pre-Training Loss Curve. (a) Default, (b) Bone Supp, (c) Bone Default, (d) New Combo.

C. *Confusion Matrices and ROC Curves*

Below, we showcase the confusion matrices and ROC curves for each augmentation strategy in both fine-tuning paradigms for SimCLR, BYOL, SwAV, and DINO. The linear evaluation confusion matrices can be found in Figure 6, Figure 8, Figure 10, Figure 12. The full model fine-tuning confusion matrices can be found in Figure 7, Figure 9, Figure 11, Figure 13. The linear evaluation ROC curves can be found in Figure 14, Figure 16, Figure 18, Figure 20. The full model fine-tuning ROC curves can be found in Figure 15, Figure 17, Figure 19, Figure 21.

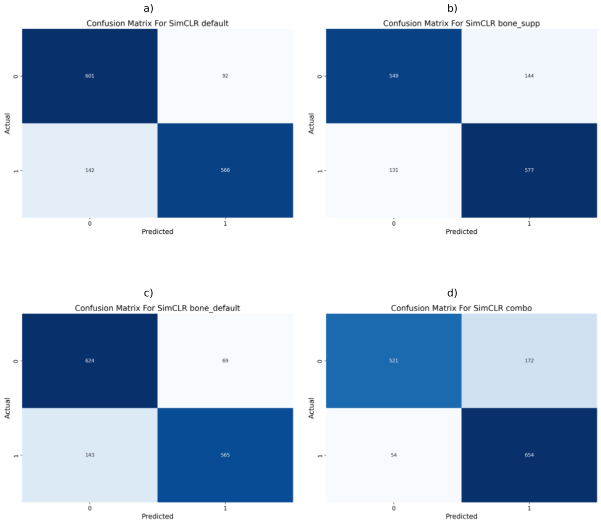

Fig. 6: SimCLR Linear Evaluation Confusion Matrices. (a) Default, (b) Bone Supp, (c) Bone Default, (d) New Combo.

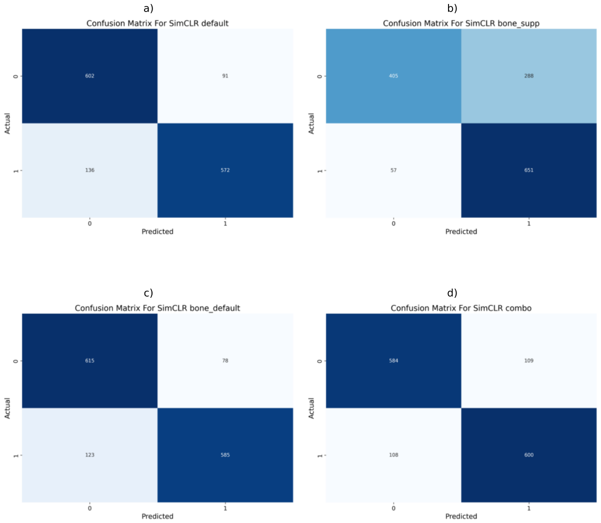

Fig. 7: SimCLR Full Model Fine-Tuning Confusion Matrices. (a) Default, (b) Bone Supp, (c) Bone Default, (d) New Combo.

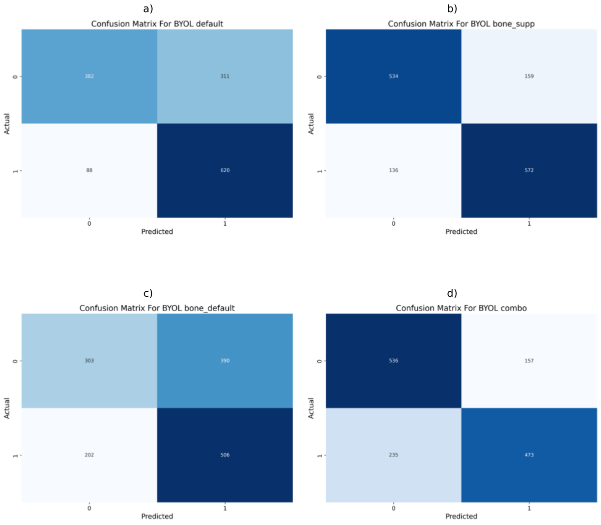

Fig. 8: BYOL Linear Evaluation Confusion Matrices. (a) Default, (b) Bone Supp, (c) Bone Default, (d) New Combo.

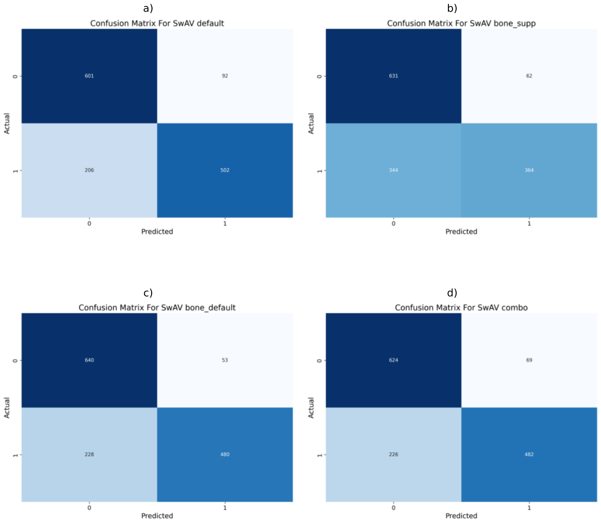

Fig. 10: SwAV Linear Evaluation Confusion Matrices. (a) Default, (b) Bone Supp, (c) Bone Default, (d) New Combo.

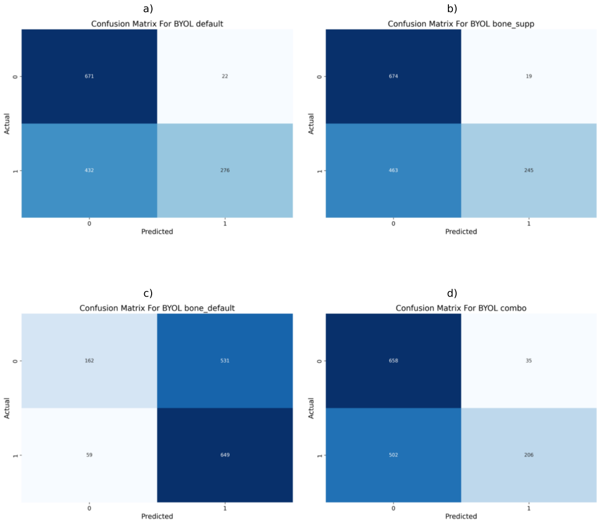

Fig. 9: BYOL Full Model Fine-Tuning Confusion Matrices. (a) Default, (b) Bone Supp, (c) Bone Default, (d) New Combo.

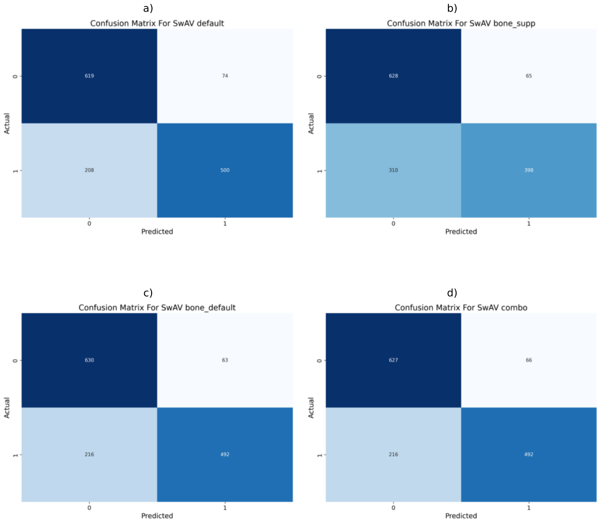

Fig. 11: SwAV Full Model Fine-Tuning Confusion Matrices. (a) Default, (b) Bone Supp, (c) Bone Default, (d) New Combo.

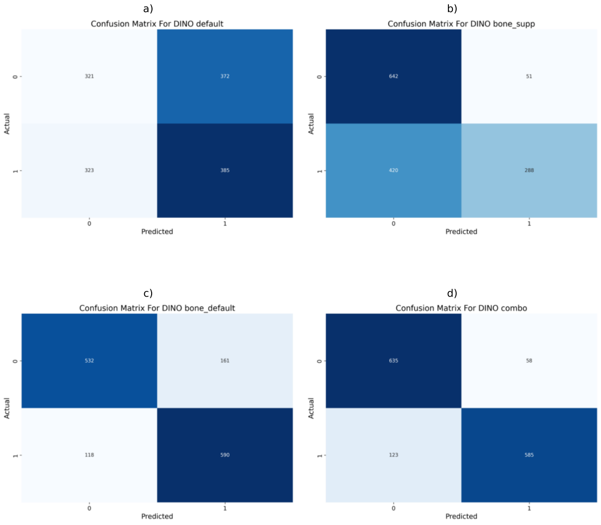

Fig. 12: DINO Linear Evaluation Confusion Matrices. (a) Default, (b) Bone Supp, (c) Bone Default, (d) New Combo.

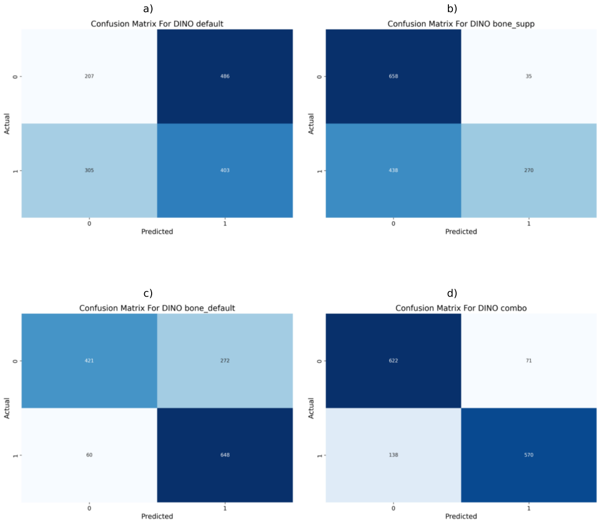

Fig. 13: DINO Full Model Fine-Tuning Confusion Matrices. (a) Default, (b) Bone Supp, (c) Bone Default, (d) New Combo.

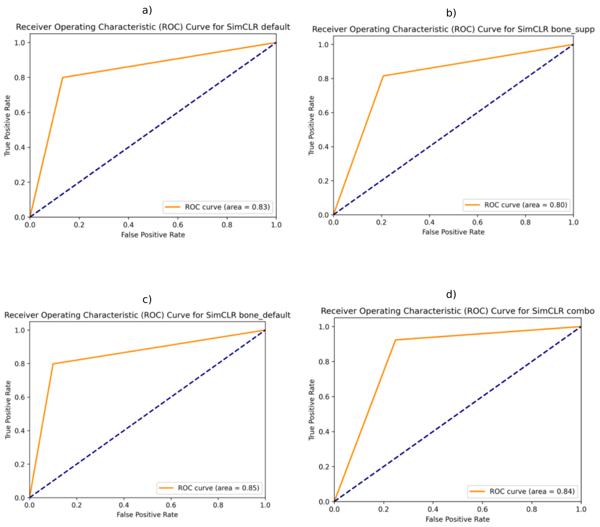

Fig. 14: SimCLR Linear Evaluation ROC Curves. (a) Default, (b) Bone Supp, (c) Bone Default, (d) New Combo.

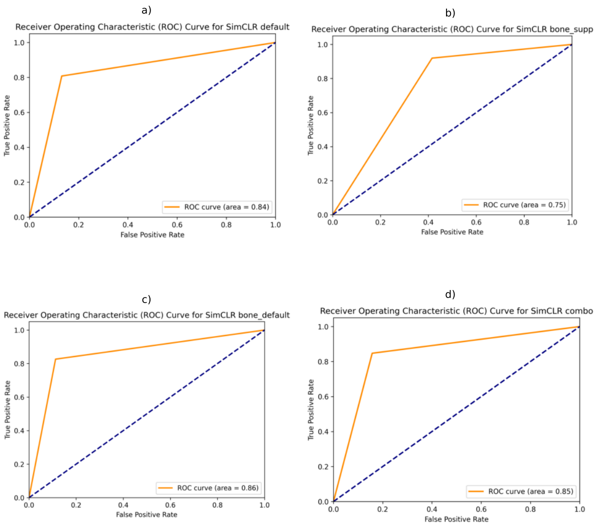

Fig. 15: SimCLR Full Model Fine-Tuning ROC Curves. (a) Default, (b) Bone Supp, (c) Bone Default, (d) New Combo.

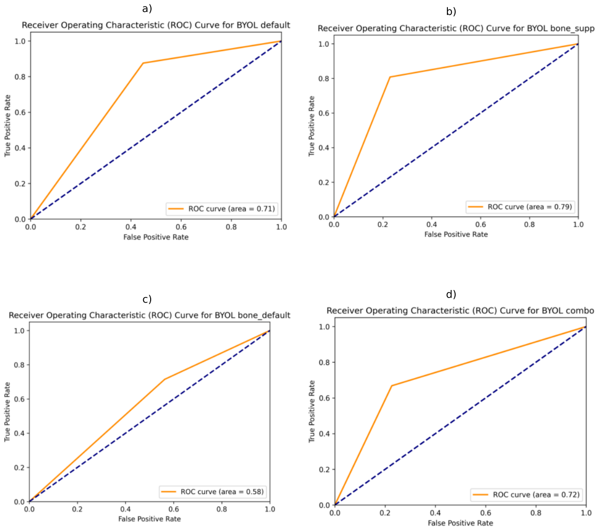

Fig. 16: BYOL Linear Evaluation ROC Curves. (a) Default, (b) Bone Supp, (c) Bone Default, (d) New Combo.

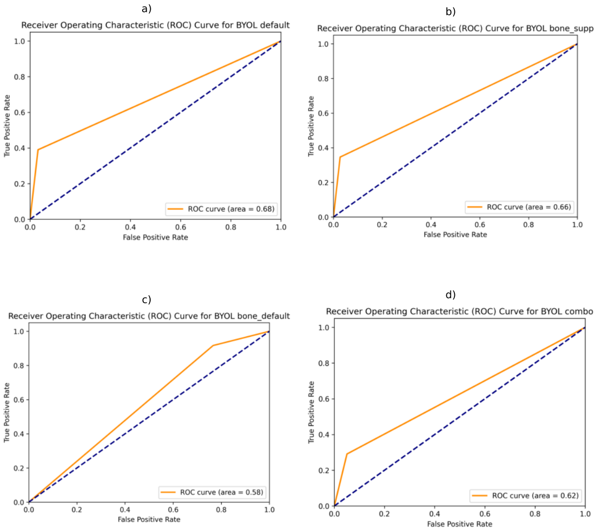

Fig. 17: BYOL Full Model Fine-Tuning ROC Curves. (a) Default, (b) Bone Supp, (c) Bone Default, (d) New Combo.

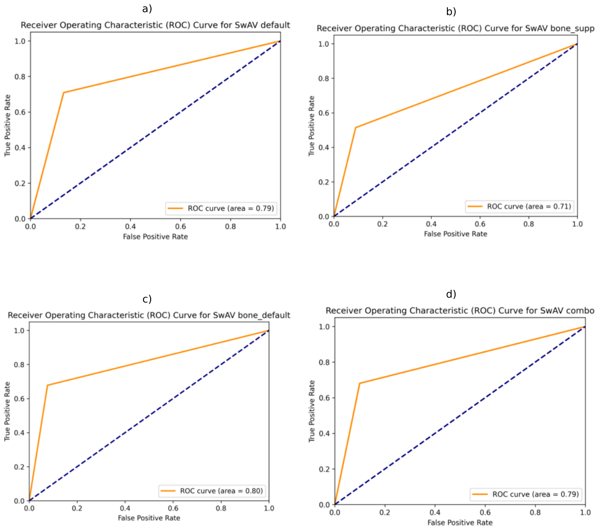

Fig. 18: SwAV Linear Evaluation ROC Curves. (a) Default, (b) Bone Supp, (c) Bone Default, (d) New Combo.

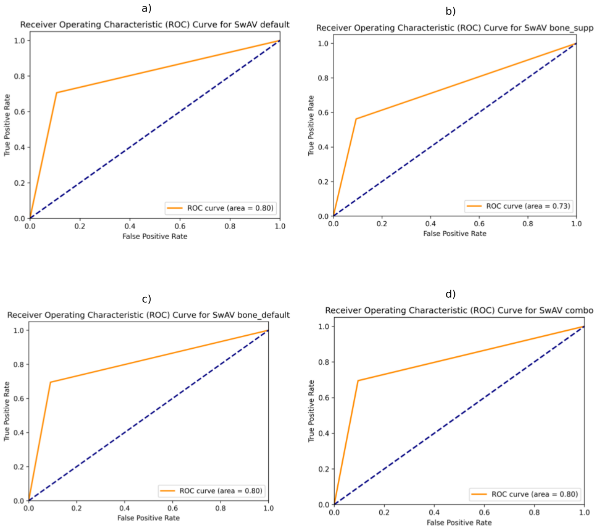

Fig. 19: SwAV Full Model Fine-Tuning ROC Curves. (a) Default, (b) Bone Supp, (c) Bone Default, (d) New Combo.

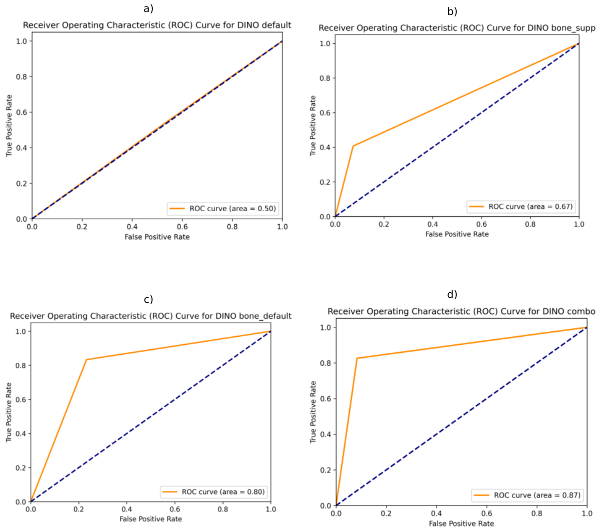

Fig. 20: DINO Linear Evaluation ROC Curves. (a) Default, (b) Bone Supp, (c) Bone Default, (d) New Combo.

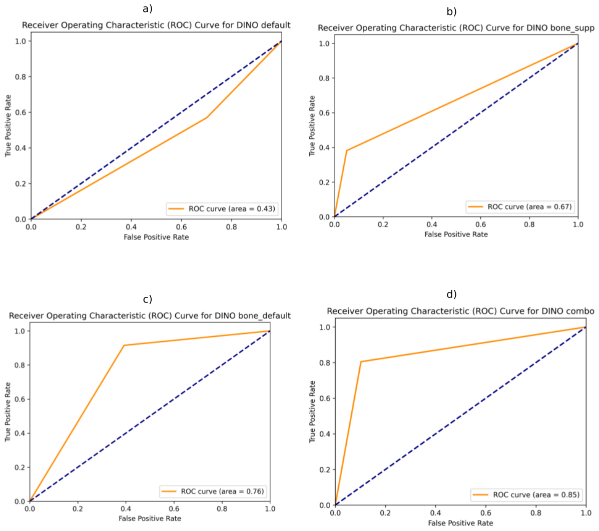

Fig. 21: DINO Full Model Fine-Tuning ROC Curves. (a) Default, (b) Bone Supp, (c) Bone Default, (d) New Combo.

References

[1] World Health Organization. *World health statistics 2022: Monitoring health for the SDGs, sustainable development goals*. May 2022.

[2] World Health Organization. *Global Tuberculosis Report 2022*. Oct. 2022.

[3] World Health Organization. “Early detection of tuberculosis: an overview of approaches, guidelines and tools”. In: (2011).

[4] Anna H van’t Hoog et al. “High sensitivity of chest radiograph reading by clinical officers in a tuberculosis prevalence survey”. In: *The International journal of tuberculosis and lung disease* 15.10 (2011), pp. 1308–1314.

[5] Karen R Steingart et al. “Fluorescence versus conventional sputum smear microscopy for tuberculosis: a systematic review”. In: *The Lancet infectious diseases* 6.9 (2006), pp. 570–581.

[6] Barbara A Styrt et al. “Turnaround times for mycobacterial cultures.” In: *Journal of Clinical Microbiology* 35.4 (1997), p. 1041.

[7] Antonina A Votintseva et al. “Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples”. In: *Journal of clinical microbiology* 55.5 (2017), pp. 1285–1298.

[8] Wim Naudé. “Artificial intelligence vs COVID-19: limitations, constraints and pitfalls”. In: *AI & society* 35 (2020), pp. 761–765.

[9] Martin J Willemink et al. “Preparing medical imaging data for machine learning”. In: *Radiology* 295.1 (2020), pp. 4–15.

[10] Xiao Liu et al. “Self-supervised learning: Generative or contrastive”. In: *IEEE Transactions on Knowledge and Data Engineering* 35.1 (2021), pp. 857–876.

[11] Sivaramakrishnan Rajaraman et al. “Chest x-ray bone suppression for improving classification of tuberculosis-consistent findings”. In: *Diagnostics* 11.5 (2021), p. 840.

[12] Ting Chen et al. “A simple framework for contrastive learning of visual representations”. In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607.

[13] Jean-Bastien Grill et al. “Bootstrap your own latent-a new approach to self-supervised learning”. In: *Advances in neural information processing systems* 33 (2020), pp. 21271–21284.

[14] Mathilde Caron et al. “Unsupervised learning of visual features by contrasting cluster assignments”. In: *Advances in neural information processing systems* 33 (2020), pp. 9912–9924.

[15] Mathilde Caron et al. “Emerging properties in self-supervised vision transformers”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 9650–9660.

[16] Kaiming He et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[17] Xu Han et al. “Pre-trained models: Past, present and future”. In: *AI Open* 2 (2021), pp. 225–250.

[18] KC Santosh et al. “Advances in Deep Learning for Tuberculosis Screening Using Chest X-Rays: The Last 5 Years Review”. In: *Journal of Medical Systems* 46.11 (2022), p. 82.

[19] Soumya Ranjan Nayak et al. “Application of deep learning techniques for detection of COVID-19 cases using chest X-ray images: A comprehensive study”. In: *Biomedical Signal Processing and Control* 64 (2021), p. 102365.

[20] Stefan Jaeger et al. “Automatic screening for tuberculosis in chest radiographs: a survey”. In: *Quantitative imaging in medicine and surgery* 3.2 (2013), p. 89.

[21] Emilio Soria Olivas et al. *Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques: Algorithms, methods, and techniques*. IGI global, 2009.

[22] Justin Ker et al. “Deep learning applications in medical image analysis”. In: *Ieee Access* 6 (2017), pp. 9375–9389.

[23] Ross Girshick et al. “Rich feature hierarchies for accurate object detection and semantic segmentation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.

[24] Maxime Oquab et al. “Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2014.

[25] Jiashu Xu. “A review of self-supervised learning methods in the field of medical image analysis”. In: *International Journal of Image, Graphics and Signal Processing (IJIGSP)* 13.4 (2021), pp. 33–46.

[26] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. “Distilling the knowledge in a neural network”. In: *arXiv preprint arXiv:1503.02531* (2015).

[27] Thomas G Dietterich. “Ensemble methods in machine learning”. In: *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings 1*. Springer. 2000, pp. 1–15.

[28] Muhammad Ayaz, Furqan Shaukat, and Gulistan Raja. “Ensemble learning based automatic detection of tuberculosis in chest X-ray images using hybrid feature descriptors”. In: *Physical and Engineering Sciences in Medicine* 44.1 (2021), pp. 183–194.

[29] Mohd Hanafi Ahmad Hijazi et al. “Ensemble deep learning for tuberculosis detection using chest X-Ray and canny edge detected images”. In: *IAES International Journal of Artificial Intelligence* 8.4 (2019), p. 429.

[30] Ekin Tiu et al. “Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning”. In: *Nature Biomedical Engineering* (2022), pp. 1–8.

[31] Mohammad Reza Hosseinzadeh Taher et al. “CAid: Context-aware instance discrimination for self-supervised learning in medical imaging”. In: *International Conference on Medical Imaging with Deep Learning*. PMLR. 2022, pp. 535–551.

[32] Matej Gazda et al. “Self-supervised deep convolutional neural network for chest x-ray classification”. In: *IEEE Access* 9 (2021), pp. 151972–151982.

[33] Mingu Kang et al. “Benchmarking Self-Supervised Learning on Diverse Pathology Datasets”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 3344–3354.

[34] Christos Matsoukas et al. “What makes transfer learning work for medical images: Feature reuse & other factors”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 9225–9234.

[35] Jia Deng et al. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[36] Yun Liu et al. “Rethinking computer-aided tuberculosis diagnosis”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 2646–2655.

[37] Arun Chauhan, Devesh Chauhan, and Chittaranjan Rout. “Role of gist and PHOG features in computer-aided diagnosis of tuberculosis without segmentation”. In: *PloS One* 9.11 (2014), e112980.

[38] Stefan Jaeger et al. “Two public chest X-ray datasets for computer-aided screening of pulmonary diseases”. In: *Quantitative Imaging in Medicine and Surgery* 4.6 (2014), p. 475.

[39] Tawsifur Rahman, Amith Khandakar, and Muhammad Enamul Hoque Chowdhury. *Tuberculosis (TB) Chest X-ray Database*. 2020. DOI: 10.21227/mps8-kb56. URL: https://dx.doi.org/10.21227/mps8-kb56.

[40] Richard Zhang, Phillip Isola, and Alexei A Efros. “Colorful image colorization”. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer. 2016, pp. 649–666.

[41] Kaiming He et al. “Momentum contrast for unsupervised visual representation learning”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9729–9738.

[42] Vinod Nair and Geoffrey E Hinton. “Rectified linear units improve restricted boltzmann machines”. In: *Proceedings of the 27th international conference on machine learning (ICML-10)*. 2010, pp. 807–814.

[43] Ilya Loshchilov and Frank Hutter. “Sgdr: Stochastic gradient descent with warm restarts”. In: *arXiv preprint arXiv:1608.03983* (2016).