applied sciences

MDPI

Article

Investigating the Performance of FixMatch for COVID-19 Detection in Chest X-rays

Ali Reza Sajun *, Imran Zualkernan and Donthi Sankalpa

Computer Science and Engineering Department, American University of Sharjah, Sharjah P.O. Box 26666, United Arab Emirates; izualkernan@aus.edu (I.Z.); g00062902@aus.edu (D.S.)
* Correspondence: b00068908@aus.edu

Featured Application: Semi-supervised learning can effectively be used to detect the chest X-rays affected by COVID-19.

Abstract: The advent of the COVID-19 pandemic has resulted in medical resources being stretched to their limits. Chest X-rays are one method of diagnosing COVID-19; they are used due to their high efficacy. However, detecting COVID-19 manually by using these images is time-consuming and expensive. While neural networks can be trained to detect COVID-19, doing so requires large amounts of labeled data, which are expensive to collect and code. One approach is to use semi-supervised neural networks to detect COVID-19 based on a very small number of labeled images. This paper explores how well such an approach could work. The FixMatch algorithm, which is a state-of-the-art semi-supervised classification algorithm, was trained on chest X-rays to detect COVID-19, Viral Pneumonia, Bacterial Pneumonia and Lung Opacity. The model was trained with decreasing levels of labeled data and compared with the best supervised CNN models, using transfer learning. FixMatch was able to achieve a COVID F1-score of 0.94 with only 80 labeled samples per class and an overall macro-average F1-score of 0.68 with only 20 labeled samples per class. Furthermore, an exploratory analysis was conducted to determine the performance of FixMatch to detect COVID-19 when trained with imbalanced data. The results show a predictable drop in performance as compared to training with uniform data; however, a statistical analysis suggests that FixMatch may be somewhat robust to data imbalance, as in many cases, and the same types of mistakes are made when the amount of labeled data is decreased.

Keywords: COVID-19; chest X-rays; deep learning; semi-supervised learning; FixMatch

Citation: Sajun, A.R.; Zualkernan, I.; Sankalpa, D. Investigating the Performance of FixMatch for COVID-19 Detection in Chest X-rays. *Appl. Sci.* **2022**, *12*, 4694. https://doi.org/10.3390/app12094694

Academic Editors: Keun Ho Ryu and Nipon Theera-Umpon

Received: 6 April 2022
Accepted: 2 May 2022
Published: 6 May 2022

1. Introduction

COVID-19 was first declared a global pandemic in March 2020 by the director of the World Health Organization (WHO) [1], and the world is still suffering from its impact. To this date, almost 425 million people have been infected worldwide, with almost 6 million deaths being recorded [2]. The disease has cold-like symptoms and is spread through droplets in the air when people cough, sneeze or even talk [3]. Although the most commonly used test to detect COVID-19 is the reverse-transcriptase polymerase chain reaction (RT-PCR) test, as it is said to have sufficient analytical sensitivity to detect the viral infection in the pre-infectious stage in an infected individual [4]; a study in Canada shows that the test had a false negative rate (FNR) of 9.3% [5]. Despite the small value FNR, considering how fast the disease spreads, the FNR could become a big issue. Hence, an alternative method of testing is required for people who show symptoms but get negative test results. One such way is the usage of chest X-rays (CXRs), as they are less costly compared to the other radiological imaging methods and have the least risk, due to low amounts of radiation.

There is considerable work applying machine learning and medical imaging techniques to reduce the burden on radiologists [6]. Due to the recent introduction of COVID-19,

automated diagnosis of COVID-19 through computer chest X-rays (CXRs) has become a popular topic [7]. As machine learning techniques required labeled data, authors such as Vantaggiato et al. [8] have provided large public datasets with a collection of COVID-19 related CXRs. However, it took almost 2 years after the start of the pandemic for such datasets to appear. This is because, in situations where a novel disease is discovered, compiling and building datasets containing adequate amounts of labeled data relating to the disease consumes considerable time and effort [9]. This is additionally taxing as the datasets generally have to be compiled by medical experts whose time would be better utilized in actively helping to contain the spread of any such diseases. Consequently, such situations, semi-supervised learning can be leveraged. As opposed to requiring large amounts of labeled data, semi-supervised algorithms are generally able to effectively generalize data classes based on only a handful of representative samples [10]. Therefore, it may be worthwhile to apply such techniques to the problem of COVID-19 detection from CXR images. This paper investigates how well the currently best performing semi-supervised learning algorithm called FixMatch by Sohn et al. [11] performs at detecting COVID-19 based on X-ray images. The contributions of this paper are as follows:

1. Providing a comprehensive literature review of using X-ray data to detect various COVID-19 types of diseases.
2. Evaluating the performance of InceptionV3, DenseNet121, Xception, ResNet50 and EfficientNet (B1, B2 and B3) to detect COVID-19 by using transfer learning.
3. Evaluating FixMatch algorithm to detect COVID-19 by using varying percentages of labeled data.
4. Evaluating the robustness of the FixMatch algorithm against data imbalance for detecting COVID-19 by varying the distribution of majority/minority classes and percentage of labeled data.

The following sections discuss various approaches to the problem of CXR classification seen in the literature before outlining the experimental methodology and discussing the results obtained.

2. Literature Review

2.1. X-ray Classification Using Deep Learning

As pointed out by Calli et al. [6], the most common work performed in the field of X-ray classification is diagnosing Pneumonia and Tuberculosis (TB).

A common dataset used for the Pneumonia problem is the ChestX-ray14 dataset that consists of over 100,000 images and 14 class labels. The cheXNet model by Rajpurkar et al. [12] that used transfer learning for diagnosis using X-rays. CheXNet is an adaptation of DenseNet [13], which was trained on the ChestX-ray14 dataset. The weights were initialized with the ImageNet weights. For pneumonia detection, the Pneumonia labeled images were taken as the positive class, and all other labels were considered as the negative class. To test the model, an unseen dataset of 420 images from four practicing radiologists at Stanford University was used. The model was also extended to classify all 14 labels and achieved state-of-the-art results at that time, with an example of 0.9164 Area Under the Curve (AUC) for detecting Hernia and 0.7680 AUC for detecting Pneumonia. This was an improvement on Wang et al. [14], who found the best model to be ResNet-50 [15] trained on a ChestX-ray8 dataset that consisted of over 32,000 unique patients with eight disease labels had an AUC of 0.63 for Pneumonia classification. Baltrushchat et al. [16] used various ResNet models and found that ResNet-38 improved the results further with an example of AUC of 0.937 for Hernia and 0.714 for Pneumonia. Irfan et al. [17] used three models for transfer learning, namely ResNet-50, Inception V3 [18] and DenseNet121. DenseNet121 performed the best, with and AUC of 0.71. The model was also tested on an unseen CheXpert dataset, and DenseNet121 performed the best, again, with an AUC of 0.76.

One of the first works to automate the detection of Tuberculosis (TB) through chest X-rays was performed by Hwang et al. [19]. The CNN architecture based on AlexNet [20], using transfer learning with low-level filters, worked the best, with an average accuracy of 90.3% across three datasets, which was a significant improvement from 77.3% when trained without transfer learning. The dataset used for training was the KIT dataset that consisted of 7020 normal and 3828 TB images retrieved from the Korean National Tuberculosis Association. Liu et al. [21] trained their model on a private dataset that consisted of six variations of TB. There was a total of 453 normal and 4248 TB images. This work also utilized AlexNet, with an accuracy of 85.86%. The model was first trained by using the original weights and then fine-tuned. They also trained a binary classifier which achieved an average accuracy of 97.82% across 5-fold evaluation. More recently, Rahman et al. [22] proposed using segmentation to extract the lung features as an input. Consequently, U-net [23] was used for the segmentation, and after trying nine different deep CNN models, it was found that ChexNet performed the best, with an accuracy of 96.47%, without segmentation; and DenseNet201 performed the best, with an accuracy of 98.6%, with segmentation, and this was an improvement from before. The dataset used to train the lung segmentation was a Kaggle chest X-ray dataset that consisted of 704 images, along with their corresponding lung masks, while NLM, Belarus, NIAID and RSNNA were used for the classification network with 3500 images of each of the binary classes.

Semi-supervised learning (SSL) has also been used to address Pneumonia detection problem using X-rays. For example, Amin et al. [24] proposed an SSL-based model utilizing Generative Adversarial Networks (GANs) [25]. GANs are used, along with transfer learning, utilizing a pretrained VGG16 model [26] as the discriminator. A few of the layers of the VGG model were unfrozen while training for fine-tuning the model toward X-ray images. The SGAN model not only generates new fake images but also learned to classify unlabeled images into the correct classes. The authors trained the model on a dataset from the Guangzhou women's and children's hospital that consisted of 5856 images, out of which 4237 were Pneumonia, and the rest were normal. For their experimentation, they randomly selected 1000 images of each class and used 70% unlabeled data for each class. The authors were able to achieve an accuracy of 94.73%, which is on par with non-SSL networks, even with unlabeled data present. Similarly, Zhang et al. [27] used unlabeled images to train a feature extractor instead of predicting pseudo labels. The authors used a c3 schema for feature extraction on unlabeled images and used Resnet50 for the final classification. The model was trained by using the ChestXRay2017 dataset that consisted of 5856 Pneumonia images and a mix of two sets of datasets from the US National Library of medicine to get 800 images of Tuberculosis. The class Pneumonia was trained with 1536 labels and achieved an AUC of 0.98, while the Tuberculosis class was trained with 448 labels and achieved an AUC of 0.923.

Tables 1 and 2 below summarizes the work reviewed for the above section.

Table 1. Summary of work using deep learning for Pneumonia classification through CXRs.

Authors	Year	Dataset	No. of Data Points	Domain	Model	Results
Wang et al. [14]	2017	ChestX-ray8	Pneumonia: 1062 Normal: 84,312	Pneumonia	ResNet-50	AUC for Pneumonia: 0.63
Rajpurkar et al. [12]	2017	ChestX-ray14	Pneumonia: 105,408 Normal: 6712	Multi-class lung diseases	CheXNet (adaptation of DenseNet)	AUC for Hernia: 0.9164 AUC for Pneumonia: 0.7680
Baltrushchat et al. [16]	2019	ChestX-ray14	Pneumonia: 1431 Normal: 110,689	Multi-class lung diseases	ResNet-38	AUC for Hernia: 0.937 AUC for Pneumonia: 0.714
Irfan et al. [17]	2020	ChestX-ray14, CheXpert (only for testing)	112,120 CXRs	Pneumonia	DenseNet121	Training AUC: 0.71 Testing AUC: 0.76
Amin et al. [24]	2020	Guangzhou women's and children's hospital dataset	Pneumonia: 1000 Normal: 1000	Pneumonia SSL	SGAN + VGG	Accuracy: 94.73%
Zhang et al. [27]	2021	ChestXRay2017	5856 CXRs	Pneumonia, Tuberculosis	C3 schema feature extractor + ResNet50	AUC for Pneumonia: 0.98 AUC for Tuberculosis: 0.923

Table 2. Summary of work using deep learning for Tuberculosis classification through CXRs.

Authors	Year	Dataset	No. of Data Points	Domain	Model	Results
Hwang et al. [19]	2016	KIT, MC, Shenzhen	Kit TB: 3828 Normal: 7020 MC TB: 58 Normal: 80 Zhen TB: 336 Normal: 326	Tuberculosis	CNN based on AlexNet	Average accuracy across three datasets: 90.3%
Liu et al. [21]	2017	Private dataset	Normal: 453 Different manifestations of TB: 4248	Multi class Tuberculosis	AlexNet	Accuracy: 85.86%
Rahman et al. [22]	2020	Kaggle dataset for segmentation, NLM, Belarus, NIAID, RSNNA for classification	TB: 3500 Normal: 3500	Tuberculosis	Unet for segmentation, DenseNet201 for classification	Accuracy: 98.6%
Zhang et al. [27]	2021	ChestXRay2017	5856 CXRs	Pneumonia, Tuberculosis	C3 schema feature extractor + ResNet50	AUC for Pneumonia: 0.98 AUC for Tuberculosis: 0.923

2.2. COVID-19 Detection Using Deep Learning

As Shah et al. [7] point out, for COVID-19 detection using medical images, the majority of previous works used transfer learning for the prediction of labels. For instance, Oh et al. [28] used FC-DenseNet103 for feature extraction of the lung and heart contour and ResNet-18 for the classification of X-rays trained on a compilation of images from various sources, namely JSRT and NLM. Their method was able to classify Bacterial Pneumonia, Tuberculosis, Viral Pneumonia, COVID-19 and others. Even with the small dataset size, the authors were able to achieve 88.9% accuracy with segmentation mask and 79.8% accuracy without mask. Another such instance is seen in the work of Mangal et al. [29], who repurposed the CheXNet model [12] that identified pneumonia to detect COVID-19. The

CheXNet model was trained by using the ChestX-ray14 dataset, and its weights were fit for classification by using chest X-rays. The authors were able to achieve an accuracy of 90.5% on the public covid-chestxray-dataset Similarly, Apostolopous et al. [30] used two self-built datasets for experimenting, using transfer learning. The authors tried VGG19, MobileNet v2, Inception, Xception and Inception ResNet v2. They found that MobileNet performed the best, even with the small dataset size, giving an accuracy of 96.78% for binary classification and 94.72% for three-class classification. Similarly, Ozturk et al. [31] used DarkNet as the basis for their model. They trained and tested on a combination of two datasets and experimented by using a binary classifier and a multiclass classifier to detect COVID-19, normal and Pneumonia. The binary classifier achieved an accuracy of 98.08%, while the three-class classifier only achieved an accuracy of 87.02%. Narin et al. [32] created three datasets from a combination of images from Kaggle and GitHub, with the first dataset consisting of 341 COVID and 2800 normal images, the second having 341 COVID and 1493 Viral Pneumonia images and the third having 341 COVID and 2772 Bacterial Pneumonia. The best model was found to be ResNet50, with an accuracy of 96.1% for dataset 1, 99.5% for dataset 2 and 99.7% for dataset 3. Khasawaneh et al. [33] used a combination of locally acquired images and a publicly available dataset for their experimentation. The authors tested a simple 2D-CNN, MobileNet and VGG-16 with different combinations of the datasets. For the 2D-CNN, the public dataset alone performed best, with an accuracy of 96.1%, while for MobileNet, the fused dataset for training and the local dataset for testing achieved an accuracy of 98.7%. Finally, the same dataset combination as MobileNet performed the best, with an accuracy of 99% with VGG-16. Luz et al. [34] used transfer learning on the latest state-of-the-art EfficientNet model with the COVIDx dataset. They were able to achieve an overall accuracy of 93.9%, with COVID-19 sensitivity of 96.8%. AbdElhamid et al. [35] extended the pretrained Xception model by adding an addition global average pooling (GAP) layer to avoid overfitting. The authors trained their model with over 7000 images from multiple Kaggle datasets for detecting three classes, COVID-19, Pneumonia and normal and were able to achieve a testing accuracy of 99.3%. Finally, Al-Shargabi et al. [36] found the best transfer learning model to be InceptionResNetv2 out of the five models they tested. As the original dataset only had 500 images, the authors decided to generate synthetic COVID-19 CXRs by using a CGAN. They were able to achieve and accuracy of 99.72%.

While the majority of works used transfer learning for identifying COVID-19 through chest X-rays, custom designed models have also been explored. One such model is COVID-Net implemented by Wang et al. [37]. They developed their own dataset COVIDx which was used by in the future. This dataset is relatively large in comparison to other datasets comprising of 358 COVID-19, 8066 normal and 5541 pneumonia chest X-rays. The model used a lightweight residual projection-expansion-projection-extension design pattern and consists of one convolution layer to expand the features into a higher dimension, three depth-wise convolutions or learning spatial characteristics, one convolutional layer for projecting back to the lower dimension and finally a single convolutional layer to get the final features. Their model achieved an average accuracy of 93%. Similarly, Arias-Lomdono et al. [38] built on this model for their experimentation. They modified the last two dense layers and added a weighted categorical cross-entropy loss for the compensation of the class imbalance. Multiple datasets are combined to create a dataset with more than 79,500 X-ray images with over 8500 COVID-19 samples. The authors experimented with three settings: directly using the raw data, using cropped images and using lung segmentation. From this experimentation, it was found that raw data performed the best, with an accuracy of 91.67%, while segmentation came close with an accuracy of 91.53%. The cropped images performed the worst, with an accuracy of 87.64%. The COV-SNET model by Hertel et al. [39] improved upon this accuracy. They used the COVIDx dataset for their experimentation. Due to the high imbalance of the dataset, the authors decided to do a 90–10 train–test split. For testing, the authors matched the number of images within the COVID-19 category. The authors also built a secondary training set consisting of 3913 COVID-19 images. The additional

images were extracted from the MIDRC-RICORD-1C dataset and BIMCV dataset. The authors improved upon the DenseNet-121 network by adding an additional dense layer with 128 units and a dropout layer, followed by a softmax layer for multiclass classification and a dense layer with a sigmoid activation function for the binary classifier. Initially, the newer layers were trained, and then the full model was unfrozen. Both the three-class and two-class classifiers achieved an average accuracy of 95%.

Finally, Sahlol et al. [40] proposed the usage of modified CNN by retraining a pretrained Inception model, along with the Fractional-Order Marine Predators Algorithm for better feature extraction. The authors built two of their own datasets by combining various sources, as well as extracting normal X-rays from Kaggle. Both the datasets had similar characteristics; for example, the age group of patients was limited to 40–84 years, and the imbalance within the classes were very high, with COVID-19 being the minority. This new approach achieved an accuracy of 98.21% on dataset 1 and 99.1% on dataset 2.

The use of ensemble networks was also experimented with to detect COVID. Kedia et al. [41] proposed using ensemble network for the classification of COVID-19 through X-rays. This method achieved an accuracy of 98.28% and was able to identify three classes being COVID-19, Pneumonia and normal. They also trained a binary classifier and was able to achieve an accuracy of 99.71%. The authors used a stacked ensemble method and used two pretrained models, namely VGG19 and DenseNet121. The final classification was performed by using a Scalar Vector Model (SVM). The authors also created their own dataset by extracting images from five different sources, including Kaggle and GitHub repositories. The final dataset consisted of 798 COVID-19 images, and this amount is larger than the number of images present in the previously reviewed works above. Similarly, Vantaggiato et al. [8] created their own dataset. Their dataset consisted of two different sets, with the first split into three categories, namely COVID-19, Pneumonia and normal, and the second split into COVID-19, Bacterial Pneumonia, Viral Pneumonia, Lung Opacity not Pneumonia and normal. This is the first dataset to split the groups into five to further distinguish between different forms of Pneumonia. For training, the authors used 404 images of each class to have a balanced set, along with 12 augmented versions of each image. The authors set aside 100 images for validation and 207 for testing. They used Argmax to get the final prediction after getting probabilities of each class from the three model within the ensemble model. For this paper, ResNet-50, Inception-V3 and DenseNet-161 were used. The authors were able to achieve an accuracy of 75.23% for the three-class dataset and 81.0% for the five-class dataset. Mahanty et al. [42] proposed the use of the Xception model, along with a Choquet Fuzzy ensemble scoring method with a balanced Kaggle dataset, and was able to achieve an accuracy of 99.57%. Win et al. [43] proposed a voting-based ensemble that consisted of the top five models from five different scenarios. These five scenarios were introduced to remove the effect of unbalanced data. Each scenario was applied on eleven different transfer learned models, such as XceptionNet, VGG and Resnet. The input to the models was lung segments from the CXRs, and the five scenarios were weighted loss, image augmentation, undersampling, oversampling and hybrid sampling. They tested their ensemble network on a combination dataset of Kaggle and GitHub and achieved a highest accuracy of 99.23%, using an ensemble of XceptionNet, MobileNetv2, DenseNet201, InceptionResNetV2 and NasNetMobile with image augmentation.

Table 3 below summarizes the work reviewed for the above section.

Table 3. Summary of work using deep learning for COVID-19 classification through X-rays.

Authors	Year	Dataset	No. of Data Points	Model	Accuracy
Oh et al. [28]	2020	JSRT, NLM	Normal: 191 Bacterial: 54 Tuberculosis: 57 Viral: 20 COVID-19: 180	FC-DenseNet103 for feature extraction + ResNet-18	88.9%
Arias-Lomdono et al. [38]	2020	COVIDx	Control: 49,983 Pneumonia: 24,114 COVID-19: 8573	Modified COVID-Net	91.67%
Wang et al. [37]	2020	Created COVIDx	COVID-19: 358 Normal: 8066 Pneumonia: 5538	COVID-Net	93%
Apostolopous et al. [30]	2020	2 Self-built datasets	Dataset_1 COVID-19: 224 Bacterial Pneumonia: 700 Normal: 504 Dataset_2 COVID-19: 224 Bacterial and viral Pneumonia: 714 Normal: 504	MobileNet	96.78%—binary classification 94.72%—3-class classification
Ozturk et al. [31]	2020	Cohen Jp + ChestX-ray8	COVID-19: 127 Pneumonia: 500 Normal: 500	DarkNet	98.08%—binary classification 87.02%—3-class classification
Sahlol et al. [40]	2020	Built own datasets	Dataset_1 COVID-19: 200 Normal: 1675 Dataset_2 COVID-19: 219 Normal: 1341	Inception + Fractional-Order Marine Predators	98.21%
Vantaggiato et al. [8]	2021	Built own datasets	3-class dataset: COVID-19: 711 + 4848 augmented Pneumonia: 711 + 4848 augmented Normal: 711 + 4848 augmented 5-class dataset: COVID-19: 711 + 4848 augmented Bacterial Pneumonia: 711 + 4848 augmented Viral Pneumonia: 711 + 4848 augmented Lung Opacity not Pneumonia: 711 + 4848 augmented Normal: 711 + 4848 augmented	Voting ensemble, ResNet-50, Inception-V3, DenseNet-161	75.23%—3-class classification 81.0%—5-class classification
Hertel et al. [39]	2021	COVIDx, MIDR-RICORD-1C, BIMCV	COVID-19: 3913 Normal: 13,417	COV-SNET (modified DenseNet-121)	95%
Luz et al. [34]	2021	COVIDx	COVID-19: 183 Pneumonia: 5521 Normal: 8066	EfficientNet	96.8%
Khasawaneh et al. [33]	2021	Local dataset + public dataset	Public dataset: COVID-19: 713 Normal: 1583 Fused dataset: 1080 Normal: 1583	MobileNet	99%
Win et al. [43]	2021	Kaggle Dataset and GitHub Dataset	COVID-19: 3616 Pneumonia: 1345 Normal: 10,192	XceptionNet, MobileNetv2, DenseNet201, InceptionResNetV2 and NasNetMobile with image augmentation	99.23%
Mahanty et al. [42]	2021	Kaggle Dataset	COVID-19: 2313 Pneumonia: 2313 Normal: 2313	Xception + Choquet Fuzzy	99.57%

Table 3. *Cont.*

Authors	Year	Dataset	No. of Data Points	Model	Accuracy
Narin et al. [32]	2021	3 datasets of combinations of Github repositories and Kaggle	Dataset_1 COVID-19: 341 Normal: 2800 Dataset_2 COVID-19: 341 Viral Pneumonia: 1493 Dataset_3 COVID-19: 341 Bacterial Pneumonia: 2772	ResNet50	Best was 99.7% for dataset 3
Kedia et al. [41]	2021	Built own datasets	3-class dataset COVID-19: 1628 Normal: 2148 Pneumonia: 2345 2-class dataset: COVID-19: 1628 Normal: 2148	Ensemble with VGG19, DenseNet121, SVM	99.71%
Al-Shargabi et al. [36]	2021	Original (500 images) + Synthetic images	Original Dataset COVID-19: 500 Pneumonia: 500 Normal: 500 Generated Images: 2790	CGAN + InceptionResNetv2	99.72%
Mangal et al. [29]	2022	COVID-chest X-ray-dataset	Normal: 1583 Bacterial Pneumonia: 2780 Viral Pneumonia: 1493 COVID-19: 155	CheXNet	90.5%
AbdElhamid et al. [35]	2022	Combination of 3 Kaggle datasets	COVID-19: 1371 Normal: 1751 Pneumonia: 4273	Xception + GAP layer + activation layer	99.3%

2.3. COVID-19 Detection Using Semi-Supervised Deep Learning

Haque et al. [44] proposed a teacher-based network where an initial CNN model was trained to generate pseudo labels for the unlabeled images, and then a classification CNN model was used to do the final classification, which was trained by using the original and pseudo labels. The authors combined three datasets, namely COVIDx, BIMCV and MIDRCRICORD, and the final dataset consisted of 3795 COVID cases, 6045 Pneumonia cases and 8851 normal cases. An equal distribution of each group was selected, and 20%, 30% and 40% labeled samples were used. From this experimentation, they found that the set with 40% labeled sampled performed the best, with an F1 score of 0.93. The 30% labeled samples were able to achieve the same F1-score of 0.91 as a fully supervised XNet model trained on the same dataset.

Some of the other works utilized a state-of-the-art SSL model named MixMatch [45] for the classification of COVID-19, with lower amounts of labeled data. For example, Calderon-Ramirez et al. [46] used MixMatch with a wide-Resnet classifier. The author used a dataset from a GitHub repository and selected 102 positive and negative images. The author then experimented on the model with 25, 40, 50, 70 and 100 labeled images. By training on the full imbalanced dataset of 4468 images, the author achieved an accuracy of 96.6%, while training with a selective number of labels saw its peak accuracy of 85.1% at 100 labeled images, as was expected, as there were only two unlabeled for each class. The authors further improved their accuracy, as seen in Reference [47], by using a Densenet121 model for the classifier with MixMatch. An additional unseen dataset was created for testing. Sets of 10, 15 and 20 labeled samples were tested with varying proportions of imbalance for the labeled set. The authors tested both binary classification and multiclass classification. The final model yielded an accuracy of 91.3% as compared to the accuracy of 67.74% for the fully supervised model. For the chest X-ray8 dataset, an accuracy of 93.4% was achieved.

Another example for the usage of SSL in the detection of COVID-19 through chest X-rays is when unlabeled images were used to train a feature extractor instead of generating pseudo labels. For example, Abbas et al. [48] proposed a three phase model for the classification of COVID-19. They compiled their own dataset with 50,000 unlabeled X-rays and a heavily imbalanced set of images including labels of COVID-19, normal, SARS and Pneumonia. First, an autoencoder was used to extract the deep local features for the unlabeled generic chest X-ray set, and then transfer learning was used to train the main model with ImageNet weights; finally, downstream training was performed to convert from generic to COVID-19 detection. They tested the model on two COVID-19 datasets, with the first dataset achieving an accuracy of 97.54% and the second dataset achieving an accuracy of 99.8%.

Gazda et al. [49] proposed a method of transfer learning in a self-supervised task-agnostic way, utilizing the CheXpert dataset for training, and then the model was tested on four datasets, namely Cell, ChestX-ray14, C19-Cohen and COVIDGR. The self-learning method is similar to the mechanism of SimCLR and MoCo, where data augmentation was used to make two versions of each image in a batch, where a positive pair is derived from the same image and a negative pair is between different images. The backbone network for this work was the ResNet50 Wide Network. The system was also tested with three different data fractions, namely 1%, 10% and 100%. For COVID detection, in the COVIDGR dataset, the 100% fraction gave the best result, 97.1% AUC; meanwhile, for the C19-Cohen dataset, the 10% fraction gave the best AUC, 91.5%.

Table 4 below summarizes the work reviewed for the above section.

Table 4. Summary of work using semi-supervised deep learning for COVID-19 classification through X-rays.

Authors	Year	Dataset	No. of Data Points	Model	Best Results
Calderon-Ramirez et al. [46]	2020	Private dataset	COVID-19: 102 Pneumonia: 102 Normal: 102	MixMatch + wide-Resnet	Accuracy: 96.6%
Calderon-Ramirez et al. [47]	2021	Private dataset	COVID-19: 102 Pneumonia: 102 Normal: 102	MixMatch + DenseNet121	Accuracy: 91.3%
Haque et al. [44]	2021	COVIDx + BIMCV + MIDRCRICORD	COVID-19: 3795 Pneumonia: 6045 Normal: 8851	CNN for pseudo label generation + CNN for classification	40% labeled sample with F1 score of 93%
Gazda et al. [49]	2021	COVIDGR, C19-Cohen	COVIDGR COVID-19: 426 Normal: 426 C19-Cohen COVID-19: 243 Normal: 564	Data augmented pairs + wide ResNet50	COVIDGR: 100% fraction: AUC of 97% C19-Cohen: 10% fraction: AUC of 91.5%
Abbas et al. [48]	2021	Private dataset	COVID-19: 576 Normal: 1583 Pneumonia: 4273	Auto Encoder + CNN with ImageNet weights	Accuracy: 99.8% on second dataset

3. Materials and Methods

3.1. The FixMatch Algorithm

The FixMatch algorithm proposed by Sohn et al. [11] in 2021 is a semi-supervised learning algorithm that makes use of the consistency regularization and pseudo-labeling techniques commonly employed as part of semi-supervised learning [9]. Consistency regularization exploits the argument that realistic perturbations of input data points should not significantly change the predictions of a model in the label space [50]. This technique is based on the manifold and smoothness assumptions required to be met by algorithms in order to effectively train in a semi-supervised manner [51]. Pseudo-labeling is the process by which the model is trained on labeled data before the unlabeled data are passed through it in order to generate predictions. These predictions are then treated as true labels to further

train the model in a supervised manner [52]. FixMatch leveraged these two techniques by a process where an artificial label is first produced by using a weakly augmented image before the model is trained to predict this label when a strongly augmented version of the same image is input to it. A loss term is then computed which consists of a supervised terms and an unsupervised term. The supervised and unsupervised loss terms can be seen in Equations (1) and (2), respectively, where p_b is the predicted distribution for the labeled sample, x_b is the true distribution (label), B is the labeled batch size, μB is the unlabeled batch size, q_b is the pseudo-label generated from the weakly augmented image and $q\theta$ is the predictions on the strongly augmented image.

$$\mathcal{L}_s = \frac{1}{B} \sum_{b=1}^{B} CE(p_b, x_b) \tag{1}$$

$$\mathcal{L}_u = \frac{1}{\mu B} \sum_{b=1}^{\mu_B} 1(\max(q_b) \geq \tau)\, CE(q_b, q\theta\ (y \mid A(u))) \tag{2}$$

This procedure is relatively simple compared to many of the prevailing state-of-the-art semi-supervised learning techniques. However, FixMatch achieved the state-of-the-art when compared with prevalent semi-supervised algorithms such as MeanTeacher [53], MixMatch [45], Unsupervised Data Augmentation [54] and ReMixMatch [55] on common benchmarking datasets such as CIFAR10, CIFAR100 [56] and SVHN [57]. The FixMatch algorithm reported an error as low as 13.81% for CIFAR-10, with just 40 labeled samples per class in contrast with the next best results of 19.10% error reported by ReMixMatch. Due to these advantages, the FixMatch algorithm was implemented as the semi-supervised learning algorithm for the work presented here.

3.2. Dataset Selection

In order to test the efficacy of semi-supervised learning in detecting diseases from chest X-ray images, a dataset compiled by Vantaggiato et al. was used [8]. The dataset consists of chest X-ray images from the 5 classes shown below:

- Normal;
- COVID-19;
- Viral Pneumonia;
- Bacterial Pneumonia;
- Lung Opacity No Pneumonia.

Figure 1 shows a sample X-ray image for each of the classes.

Each of the classes contained 404 training images and 207 testing images, resulting in a total of 2020 training images and 1035 testing images. The images making up the testing set were obtained from sources different to those used to compile the training set. The overall dataset was compiled by Vantaggiato et al. [8] and involved combining various open-source datasets, namely the IEEE8023 COVID-19 chest X-ray dataset [58], Chest X-ray Images Pneumonia [59], RSNA Pneumonia Detection Challenge [60], CheXpert [12] and the China CXR set and Montgomery set [61]. Additionally, the authors reported collecting the test images for COVID-19 from a hospital in Algeria. Given the range of diseases contained within the dataset taken from various commonly used datasets, this dataset was chosen in order to benchmark the efficacy of semi-supervised learning in classifying between the different diseases based on X-ray images.

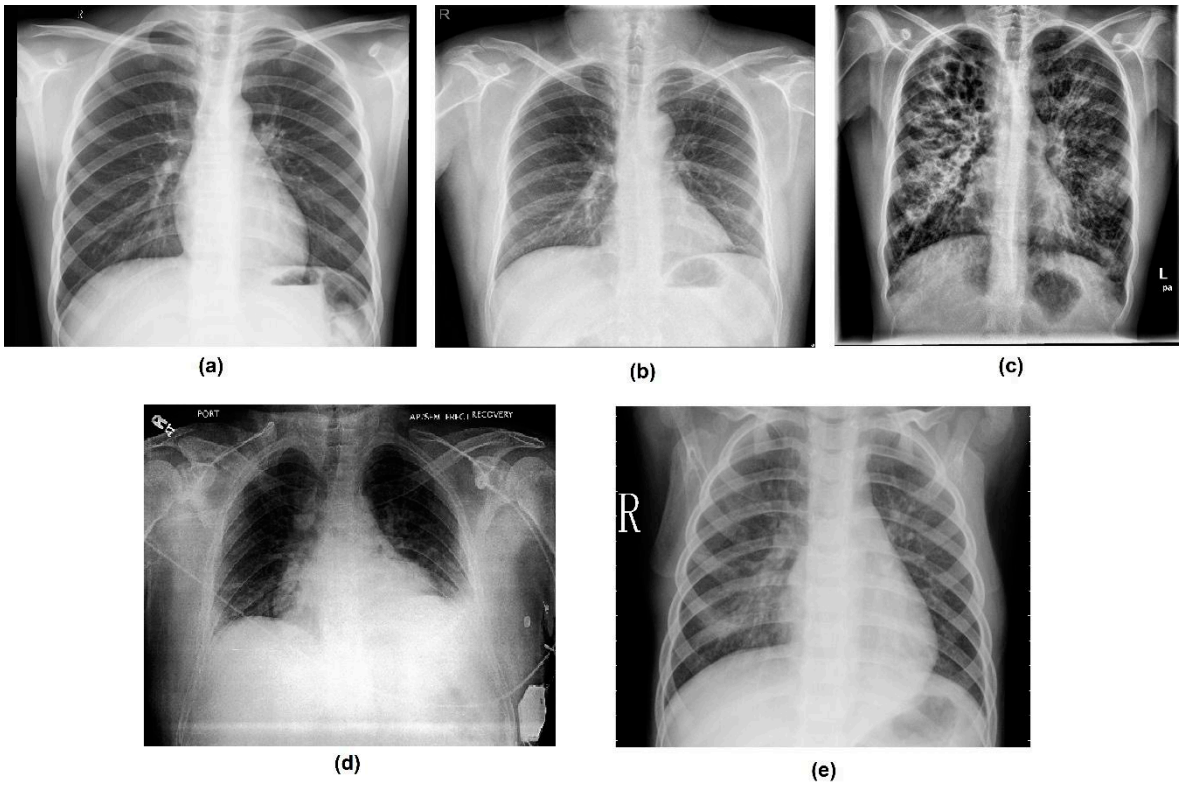

Figure 1. Sample x-ray images of (**a**) normal, (**b**) Bacterial Pneumonia, (**c**) COVID-19, (**d**) Lung Opacity and (**e**) Viral Pneumonia.

3.3. Preparing the Dataset

In order to prepare the dataset for the training, a stratified split was conducted to assign a proportion of the imbalanced data to serve as the unlabeled data for the semi-supervised learning process, while the rest would serve as the labeled data. The proportion of labeled data was varied, with the tests being conducted with the labeled proportion being 80%, 60%, 40%, 20%, 10% and, finally, an extreme case of 5%. Finally, the images were resized to 224 × 224, as this is the standard input for the underlying ResNet-18 model [15] used within the FixMatch algorithm.

3.4. Model Parameters and Training

In order to perform training of the FixMatch semi-supervised algorithm, a Pytorch [62] implementation of the algorithm [63] was used, with appropriate modifications being made to apply it to the chest X-ray dataset, as well as performing the imbalancing step. The hyperparameters used for the training were kept similar to the parameters reported to be used by the authors of FixMatch [11] for their experimentation with the ImageNet dataset [64]. Table 5 summarizes the hyperparameters used in the training process.

Table 5. Training parameters.

Parameter	Value
Batch Size	64
Epochs	300
Unlabeled Batch Size Coefficient	5
Unlabeled Loss Coefficient	10
Pseudo-label Threshold	0.7
Model	ResNet-18
Weight Decay	0.0003
Initial Learning Rate	0.4

3.5. Supervised Baseline

In order to establish a supervised baseline to compare the results of the semi-supervised algorithm, a number of prevalent CNN models were trained on the full training dataset and evaluated by using the testing subset. The models implemented were InceptionV3 [18], Xception [65], DenseNet121 [13], MobileNetV2 [66], EfficientNetB1, EfficientNetB2 and EfficientNetB3 [67]. The parameters used for this training can be seen in Table 6.

Table 6. Semi-supervised training parameters.

Parameter	Value
Batch Size	64
Epochs	100
Optimizer	Stochastic Gradient Descent
Batch Size	32
Learning Rate	0.01

4. Results and Discussion

Comparing FixMatch to Supervised Models

In order to effectively compare the performance of the FixMatch semi-supervised algorithm to existing supervised CNN approaches, the macro-average F1-scores of each of the models were computed and plotted. The formula for the F1-score can be seen in Equation (3), while the formulas for precision and recall can be seen in Equations (4) and (5).

$$\mathrm{F1-Score} = 2 \times \frac{(precision \times recall)}{(\ precions + recall)} \tag{3}$$

where we have the following:

$$Precision = \frac{True\ Positives}{True\ Positive + False\ Postive} \tag{4}$$

$$Recall = \frac{True\ Positives}{True\ Positive + False\ Negatvie} \tag{5}$$

Computing the macro-average F1-score involves computing the F1-scores for each class and then averaging it out such that each class has an equal weight in the overall score. To ensure an equal say from each class, this metric was used in order to compare the F1-scores. The class-wise F1-scores of the FixMatch model with varying amounts of labeled data can be seen in Table 7. As Table 7 shows, the average macro-average F1-score across all tests was about 0.74. However, COVID-19 was detected (from others) with a high rate of over 0.87 and as high as 0.94 and 0.93, using only 5% of the labeled data. Interestingly, the model had the most problem detecting normal against others. This is especially noteworthy because the model is trained on equal amounts of data across the classes, and, therefore, in order to gain a better understanding of the considerable difference in scores, an analysis of the data themselves is carried out in Section 4.

Table 7. Class-wise F1-score of FixMatch.

Class	FixMatch 80% Labeled	FixMatch 60% Labeled	FixMatch 40% Labeled	FixMatch 20% Labeled	FixMatch 10% Labeled	FixMatch 5% Labeled
Bacterial	0.68	0.64	0.7	0.66	0.66	0.67
COVID-19	**0.99** *	**0.97**	**0.97**	**0.94**	**0.87**	**0.93**
Lung Opacity	0.79	0.8	0.8	0.76	0.76	0.79
Normal	0.67	0.59	0.68	0.5	0.55	0.38
Viral	0.74	0.69	0.74	0.74	0.74	0.64
Macro-Average	0.77	0.74	0.78	0.72	0.71	0.68

* Bold depicts the highest F1-score.

In order to further investigate these results, the precision vs. recall graphs of the tests are shown in Figure 2.

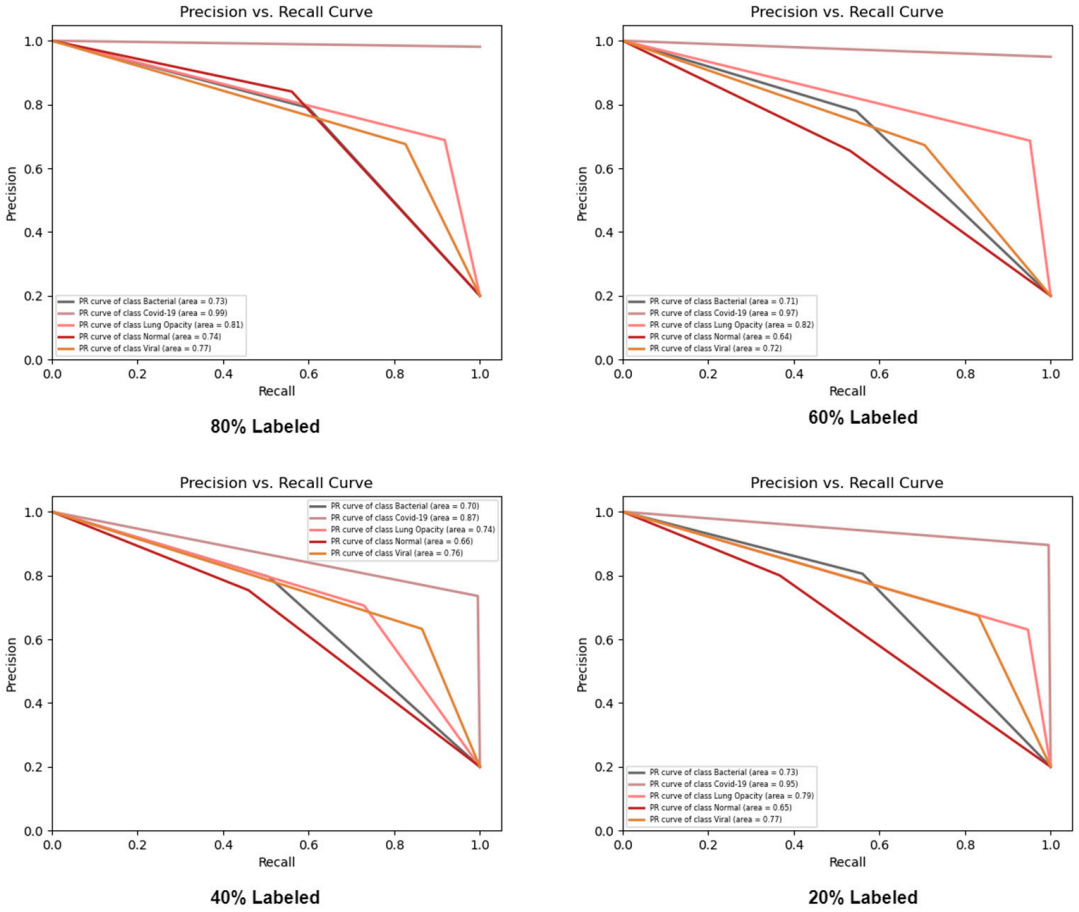

Figure 2. Precision vs. recall curves.

As the figure shows, in all cases, the COVID-19 class outperformed the remaining classes by considerable margins. As expected, the curves do become steeper as the amount of labeled data is reduced, thereby serving as an indicator for the model's deteriorated performance. It can be noted here that the recall of the COVID-19 class consistently remains high, and this is indicative of a low false negative rate. This is especially important because it means that, while there may be cases where the model may diagnose a healthy individual with COVID-19, it is less likely to fail diagnosing a patient with COVID-19. Considering how contagious the COVID-19 disease is, this is an important characteristic for the model to have and, therefore, proves the model's suitability in being used for COVID-19 detection.

Table 8 shows the results of using transfer learning with the various pretrained models. As Table 8 shows, using the complete data, InceptionV3 and DenseNet121 performed well, with F1-scores of 0.98 and 0.97. Other models, such as EfficientNet, B3 did not perform well at all.

Our analysis of the results revealed a number of interesting trends. The FixMatch algorithm is seen to be competitive to the fully supervised models even at lower amounts of labeled data. Indeed, the model was noted to perform adequately well until trained with only 20% of labeled data (corresponding to only about 80 labeled images per class), after which a drop in performance is seen, as the labeled data are reduced to 5%. It is seen that, even with just 5% of labeled data, the model performed fairly well, reporting an F1-score of 0.68, which is just 0.02 less than that reported by the fully supervised MobileNetV2 model.

Table 8. Class-wise F1-score for supervised CNN models with entire labeled dataset.

Class	InceptionV3	DenseNet121	Xception	ResNet50	MobileNetV2	EfficientNet B1	EfficientNet B2	EfficientNet B3
Bacterial	0.69	0.68	0.63	0.63	0.46	0	0.01	0
COVID-19	**0.98** *	**0.97**	**0.93**	**0.89**	**0.97**	**0.33**	**0.29**	**0.39**
Lung Opacity	0.79	0.8	0.75	0.7	0.74	0	0.51	0.04
Normal	0.61	0.63	0.59	0.68	0.63	0	0.19	0.14
Viral	0.78	0.76	0.74	0.74	0.7	0	0.5	0.6
Macro-Average	0.77	0.77	0.73	0.73	0.70	0.07	0.30	0.23

* Bold depicts the highest F1-score.

Since our class of interest was that of COVID-19, it is useful to examine the results obtained specifically for that class. In general, it is seen that the COVID-19 class performed exceptionally well as compared to the other classes. Indeed, the FixMatch algorithm has outperformed the best supervised CNN with 80% labeled data. Furthermore, it is seen that, even with 5% of labeled data, the FixMatch model was able to show an F1-score of 0.93, matching that of the Xception CNN model and outperforming the ResNet50 model. These satisfactory results with the COVID-19 class further reinforce the usability of semi-supervised learning in the domain, as well as the adaptability of using this technique in cases where labeled data are limited.

It is noteworthy that the architecture of the EfficientNet, which has lately emerged as one of the best performing CNN models, seemed to have completely failed to learn the intricacies of detecting the diseases in the chest X-rays. Three different architectures of increasing size, namely B1, B2 and B3, were implemented, but all of them gave an inadequate performance.

In order to be able to compare the performance of our models to the initial work performed by Vantaggiato et al. [8], the accuracies obtained from each of the runs were tabulated for the benchmark score of 81% obtained by the authors. The results can be seen in Table 9.

Table 9. Accuracies across the various models.

Model	Accuracy
Ensemble Approach [8]	81%
InceptionV3	78
DenseNet121	74
Xception	78
ResNet-50	74
MobileNetV2	72
FixMatch with 80% Labeled Data	78
FixMatch with 60% Labeled Data	78
FixMatch with 40% Labeled Data	78
FixMatch with 20% Labeled Data	78
FixMatch with 10% Labeled Data	73
FixMatch with 5% Labeled Data	72

It can be seen from Table 9 that, while FixMatch did outperform the CNN approaches, it remained slightly (<3%) inferior to the ensemble approach adopted by the authors, with an accuracy of just over 78% as compared to the 81% obtained by Vantaggiato et al. [8]. However, given that the underlying architecture of the FixMatch is a simple ResNet18 CNN model, we see that it is much less computationally intensive as opposed to the multi-model ensemble approach proposed by the authors.

Having established the efficacy of semi-supervised learning in the domain of COVID-19 detection by using chest X-rays, a number of tests were conducted in order to measure the impact of an imbalanced dataset on the training of the FixMatch model.

5. Exploratory Imbalance Analysis

5.1. Methodology

While the dataset selected has a balanced distribution across all classes, it may be worth investigating the performance for the semi-supervised algorithm in cases where the dataset is imbalanced in nature. This is especially relevant to the domain in question, where newer diseases would have significantly lesser available data as compared to previously established diseases. Therefore, in circumstances such as these, any detection model would be trained on various forms on imbalance data. In order to develop an imbalancing procedure, we first identified three target imbalance distributions which could simulate a variety of imbalance situations. A Weibull distribution was used in order to select the target imbalance distributions. This is a widely used model in the field of modern statistics, due to its ability to fit data from a large range of applications, from economics to engineering [68]. While a variety of variations of the Weibull distribution exist, the implementation followed by this work is based on the Weibull distribution calculator developed by Matt Bognar at the University of Iowa [69]. The probability density function that was applied in this case is seen in Equation (6).

$$f(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} e^{-(\frac{x}{\beta})^{\alpha}} \tag{6}$$

where $x > 0$, shape $\alpha > 0$ and scale $\beta > 0$.

In order to generate different distributions, the shape and scale are varied, and three of the resulting distributions are selected based on varying entropies. Figure 3 displays the final selected imbalance distributions, along with the uniform distribution, which is used as a baseline. The x-axis represents the classes, and the y-axis represents the proportion of the overall data the corresponding class occupies.

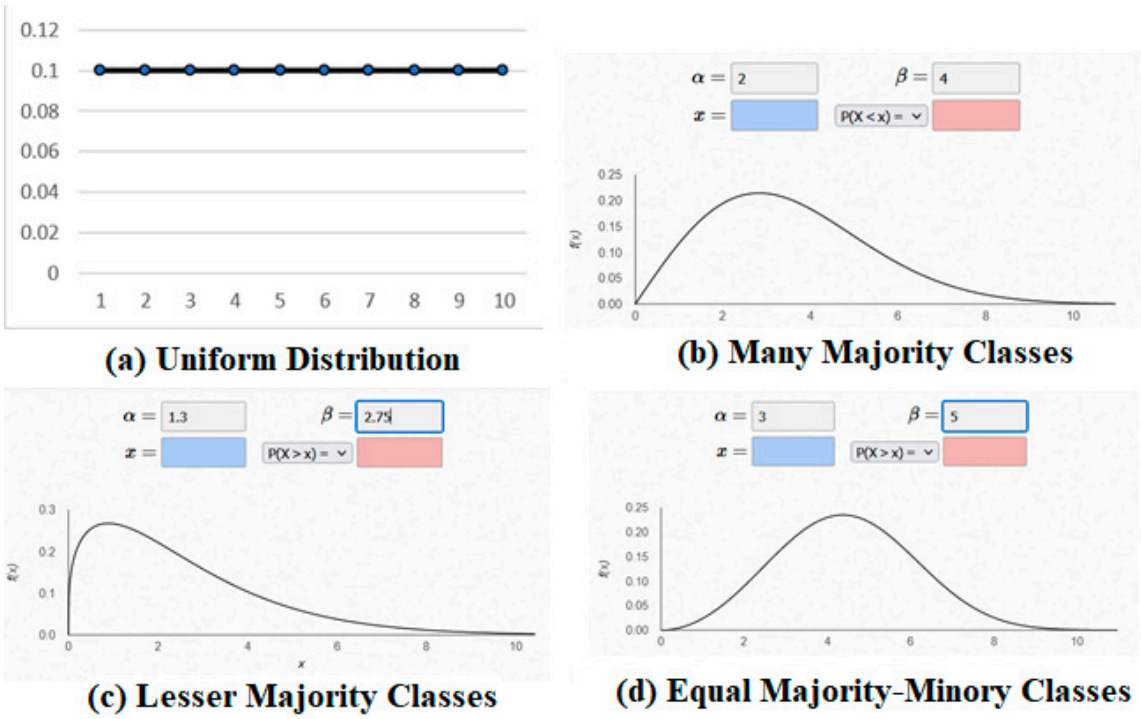

Figure 3. Various imbalance distributions used.

As can be seen in the figure, the distribution with many majority classes has the lowest number of minority classes, as its right tail is less defined than the other chosen distributions. The distribution with less majority classes has a long right tail with a well-defined peak on the left side, resulting in a very small number of samples existing in the classes corresponding to the probability on the right side. Furthermore, this long

tail leads to there being a greater number of minority classes, and this would affect the training. Finally, the equal majority–minority distribution has both a right tail and a left tail. This means that there are minority classes on both ends of the distribution, with an overwhelming majority class corresponding to the middle of the distribution. Due to the tails on both side, this leads to a larger number of minority classes, such that the proportion of majority to minority classes is approximately equal.

In order to better describe the chosen distributions, the entropies of each of the distributions are calculated based on the formula seen in Equation (7), where in $P(x_i)$ is the probability of the given point.

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i) \tag{7}$$

The resulting entropies can be seen in Table 10.

Table 10. Entropy of chosen distributions.

Distribution	Entropy
Uniform Distribution	2.3025
Many Majority	1.9305
Lesser Majority	1.9665
Equal Distribution	1.8886

To estimate the difference between the chosen distributions, the Kullback–Leibler divergence [70] and the Jensen–Shannon distance [71] between the distributions were measured. The formula used to calculate the Kullback–Leibler divergence can be seen in Equation (8), and the formula used to calculate the Jensen–Shannon distance can be seen in Equation (9), where P and Q are the two distributions being compared.

$$D_{KL}(P||Q) = \sum_{i=1}^{n} P(x_i) \log\left(\frac{P(x_i)}{Q(x_i)}\right) \tag{8}$$

$$JS(P||Q) = \frac{1}{2}KL(P||M) + \frac{1}{2}KL(Q||M) \text{ where } M = \frac{1}{2}(P+Q) \tag{9}$$

Table 11 displays the KL divergences between the distributions, while Table 12 displays the JS distance. The distributions are ascendingly ordered based on their entropy.

Table 11. KL divergence between the chosen distributions.

	Equal Distribution	Many Majority	Lesser Majority	Uniform Distribution
Equal Distribution	0	0.1661	0.6532	0.4140
Many Majority	0.2054	0	0.2383	0.3720
Lesser Majority	1.3419	0.7906	0	0.3360
Uniform Distribution	1.2642	1.2073	0.4092	0

Table 12. JS distance between the chosen distributions.

	Equal Distribution	Many Majority	Lesser Majority	Uniform Distribution
Equal Distribution	0	0.2101	0.4124	0.3440
Many Majority	0.2101	0	0.2590	0.3287
Lesser Majority	0.4124	0.2590	0	0.2940
Uniform Distribution	0.3440	0.3287	0.2940	0

As the tables illustrate, there is a significant distance between the chosen distributions, thus indicating that a wide range of imbalance situations is being considered in order to fully gain an understanding of the effect of imbalance on the training of semi-supervised learning algorithms. In order to prepare the dataset to follow the chosen imbalancing distributions, an imbalancing procedure was carried out on the chest X-ray dataset. The imbalance distributions were applied to the dataset with a randomizing process in place to assign classes to the different class proportions. The imbalancing was performed by randomly oversampling or undersampling classes based on the class weight assigned from the imbalance distribution.

5.2. Results and Discussion

5.2.1. Overall Results

The top-1 accuracies on the test data were computed for each of the tests, where the proportion of labeled samples was varied from 80% to 5% across the four levels of data imbalance. The results can be seen in Figure 4.

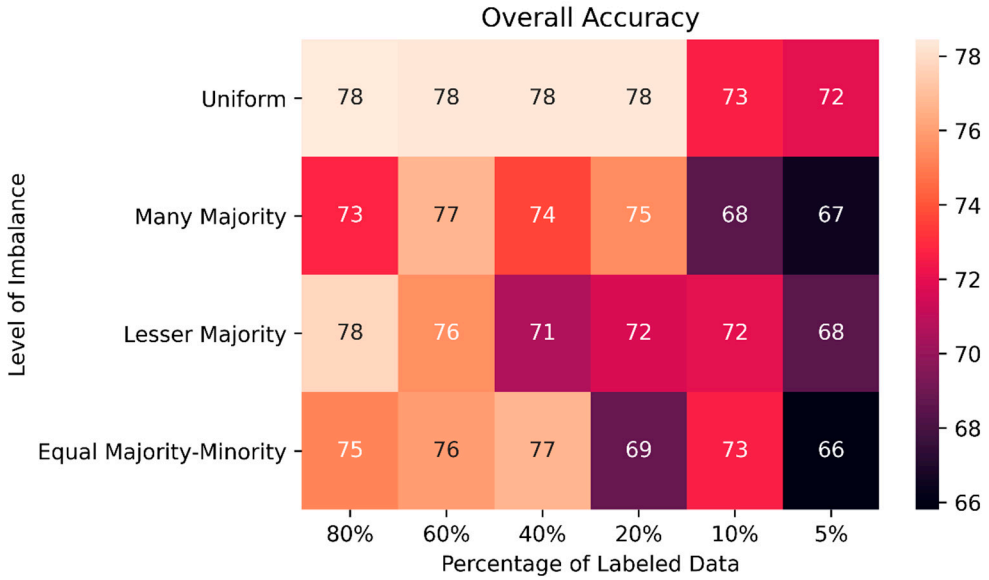

Figure 4. Overall top-1 accuracy of FixMatch for varied labeled samples and 4 levels of imbalance.

A number of interesting trends can be observed from this set of results. Firstly, as discussed previously, given uniform balance in the data, an accuracy of 78% can be obtained with just 80% labeled data. This is closely comparable to the 81% reported by Vantaggiato et al. [8], using their ensemble approach on the same dataset. A second interesting observation is that, even with 5% labeled data, the accuracy for the uniform imbalance is as high as 72%, which is only a 6% deterioration as compared with 80% labeled data and a 9% deterioration compared with fully supervised techniques. The performance with a uniform data imbalance is generally consistent with decreasing amounts of labeled data, reporting only a 1% drop until as low as 10% of the data used being labeled. This could provide great utility in the case of future outbreaks of contagious diseases, where, at the early stages, there is generally a shortage of labeled data, and medical practitioners who label the data are otherwise occupied in containing the outbreak. In order to further analyze these results, the overall macro-average F1-scores across the tests are computed as seen in Figure 5.

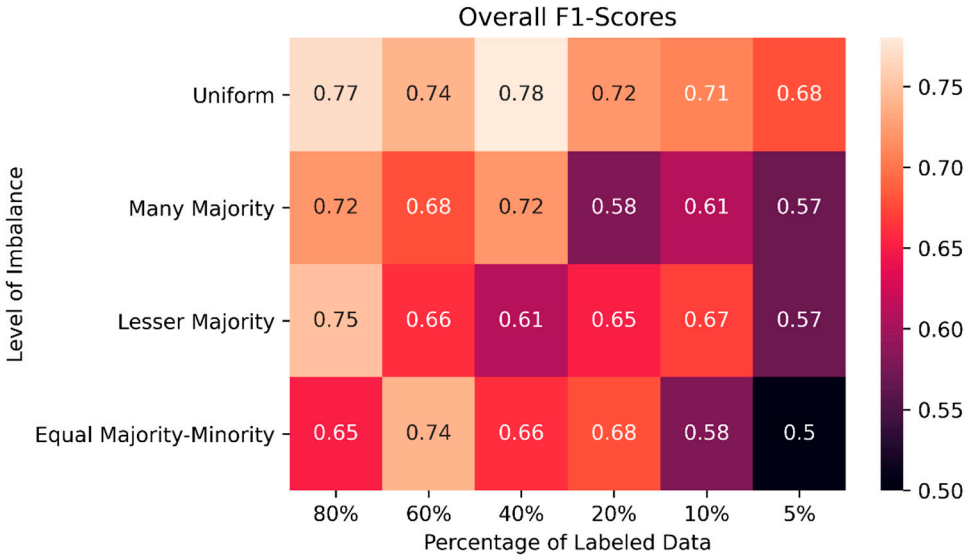

Figure 5. Overall F1-score for varied labeled samples and 4 levels of imbalance.

A greater amount of variation can be seen in the case of the F1-scores, where a greater deterioration can be seen even in the case of the uniform class. As can be seen from both result metrics, an expected drop in performance is seen when the algorithm is trained with imbalanced data. This is rather concerning, as, in the case of a newly discovered disease, samples from the disease would be much fewer than those of other similar diseases, and this would bias the model toward the majority when, indeed, the disease of interest might be the minority. Based on the results, it can be noted that the distribution containing equal numbers of majority and minority classes seemed to display the poorest performance with lower proportions of labeled data. This could particularly be due to there being well-defined majority classes which the model would tend to be biased toward, along with an equal number of minority classes which the model fails to generalize to. Similarly, the distribution with lesser majority has a long right tail leading to it that has a large number of minority classes, as is evidenced from the results. The distribution with many majority classes, while imbalanced, has a less strongly defined tail, meaning the minority classes are represented less poorly than those in the previous distributions. This is reflected in the results, as well, when looking at the accuracies and F1-scores, which reveal a general trend of the distribution, with many majority classes having performance closest to the results with the uniform balance. In general, a noted trend is seen where larger amounts of labeled data are needed by the model to learn accurately in cases where an imbalance in data is present. Indeed, the case of 80% labeled data shows the least variability in results even with a high imbalance in data, as the accuracy varies from 78% to 73%, and the F1-score varies from 0.77 to 0.65. This is in contrast to the extreme case of only 5% of the data being labeled, where the accuracy varies from 72% to 66%, and the F1-score, in particular, has a large range, as it varies from 0.68 to 0.50.

5.2.2. Class-Wise Analysis

In order to analyze the performance of the trained models further, the class-wise F1-scores of the various tests were computed and are displayed in Figures 6 and 7.

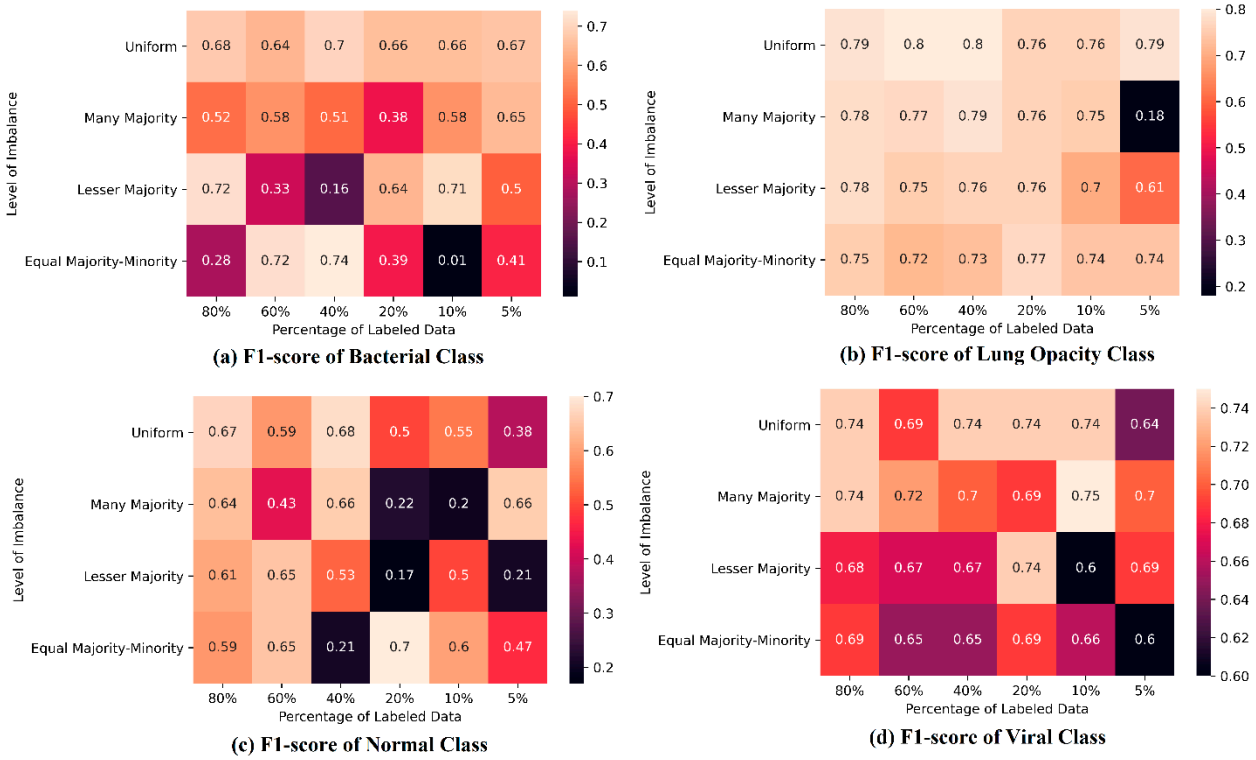

Figure 6. Class wise F1-score for varied labeled samples and 4 levels of imbalance.

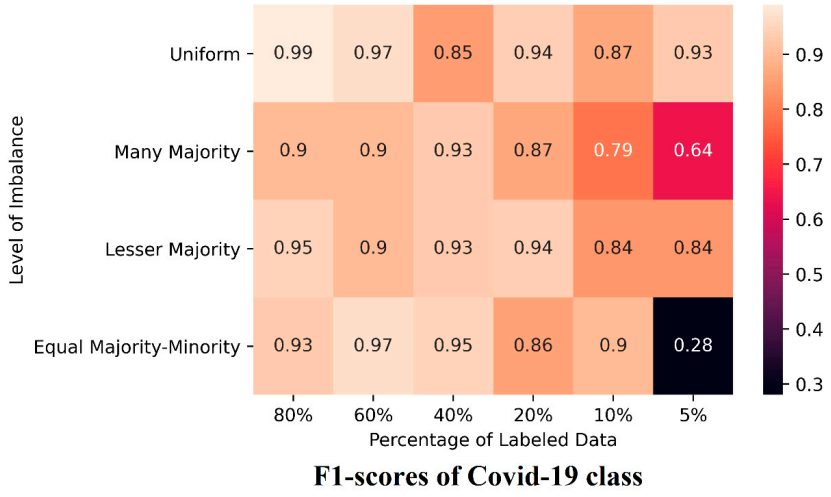

Figure 7. F1-scores of COVID-19 class for varied labeled samples and 4 levels of imbalance.

A number of interesting trends were revealed when examining this data. Firstly, we noticed that the COVID-19 class consistently outperforms the other classes and is generally seen to be extremely distinguishable by all the models, except in the extreme cases of 5% labeled data. This is rather promising, considering that the COVID-19 class is the class of interest in this case, given the recent outbreak of the pandemic. Indeed, a near-perfect score of 0.99 is seen when uniform data are used to train the model with 80% labeled data. This score remains rather consistent, even with lower proportions of labeled data, indeed, remaining as high as 0.93 in the case of 5% labeled data for the uniform distribution. While an expected drop in performance is noted for the different imbalance distributions, the overall F1-score generally remains high, thus confirming the viability of the process in detecting X-rays of persons infected with COVID-19. Such consistency in results across all imbalance distributions and percentages of labeled data is also noted in the case of the

lung opacity class, where F1-scores in the range of 0.7–0.8 are consistent across all tests. However, in the cases of the bacterial, normal and viral classes, a much greater amount of variability is observed in across the different distributions and proportions of labeled data. In order to understand these results, the PCA [72] and TSNE [73] dimensionality reduction techniques were applied to the dataset, and the results were plotted. Figure 8 shows the PCA results, and Figure 9 shows the TSNE results.

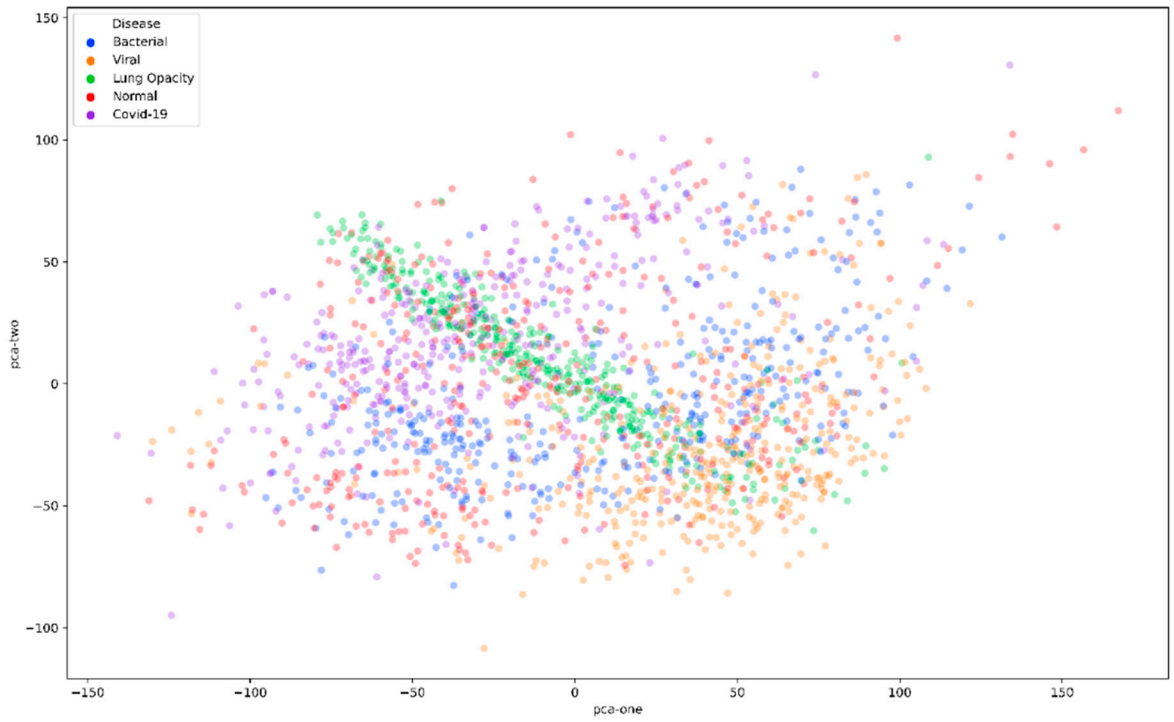

Figure 8. PCA results.

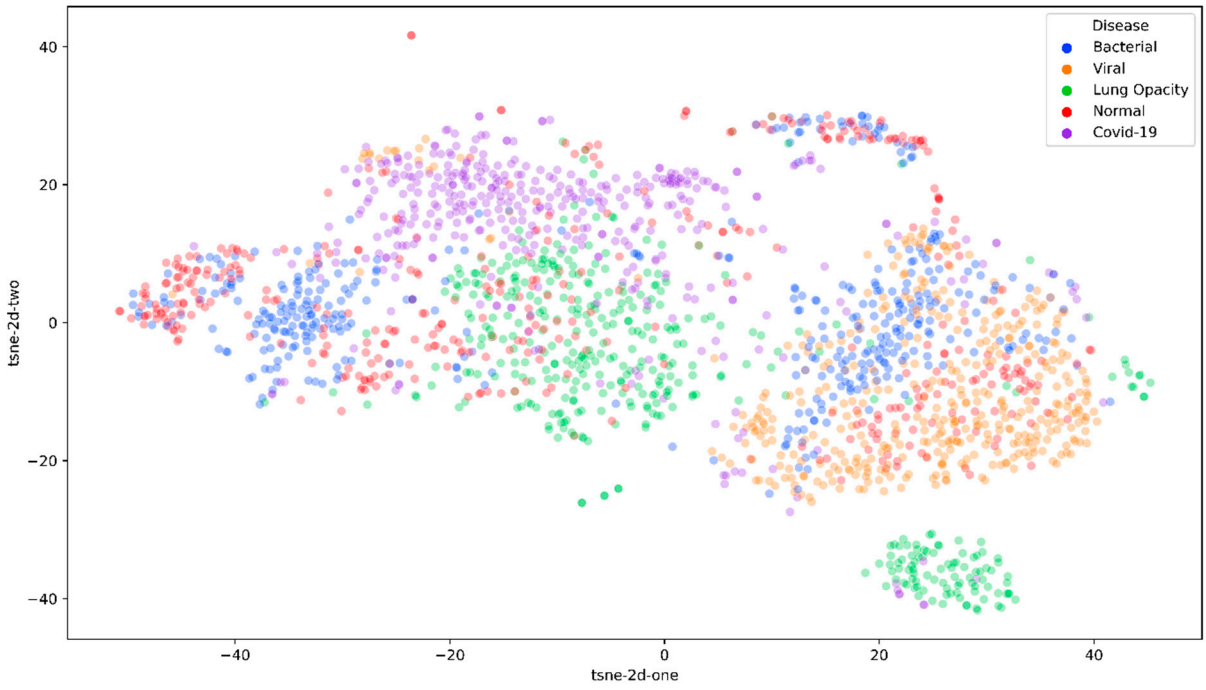

Figure 9. TSNE results.

This visualization reveals a number of interesting patterns. Firstly, the lung opacity class has a well-defined cluster in both the visualization types. It is perhaps due to this that the consistency of the F1-scores for this class is maintained despite the varying levels of imbalance and amounts of labeled data. Similarly, the COVID-19 class has a well-defined high-density cluster in the TSNE representation surrounded by areas of relatively low density, and this would clearly push the decision boundary around the COVID cluster, thereby accounting for the consistency and efficiency seen when identifying the COVID-19 class. A look at the TSNE representation reveals a high blend in the Bacterial and Viral classes that might be a factor in their failure to perform, as well as for the lower number of labeled samples. In both representations, the normal class is seen to have no defined cluster, with its data points scattered all over the representative space, therefore justifying its poor performance with respect to the other classes. Given these trends, it might be noteworthy to examine the variability in confusion matrices across the various imbalance distributions, as well as various amounts of unlabeled data in order to determine whether the same kinds of errors are made in each case.

5.2.3. Statistical Analysis of Confusion Matrices

In order to analyze the consistency in the types of errors made by the models across the various test cases, the Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMU) scores across the confusion matrices of the models were computed to determine how closely their errors matched. The formulas for these metrics are shown in Equations (10) and (11), respectively.

$$ARI = \frac{\Sigma_{ij}\binom{n_{ij}}{2} - \frac{\Sigma_i\binom{a_i}{2}\Sigma_j\binom{b_j}{2}}{\binom{n}{2}}}{\frac{\Sigma_i\binom{a_i}{2}+\Sigma_j\binom{b_j}{2}}{2} - \frac{\Sigma_i\binom{a_i}{2}\Sigma_j\binom{b_j}{2}}{\binom{n}{2}}} \tag{10}$$

where n_{ij} is the value of an element in row i and column j in the contingency table, a_i is the sum of the elements in row i, b_j is the sum of the elements in column j and n is the total count in the table.

$$AMU(U,V) = \frac{MI(U,V) - E(MI(U,V))}{avg(H(U),\ H(V)) - E(MI(U,V))} \tag{11}$$

where $MI(U,V)$ is the mutual information between U and V, $E(MI(U,V))$ is the expected mutual information between U and V, and $H(U)$ is the entropy associated with U.

The ARI and AMU were computed for various cases in order to gain a better understanding of the obtained results. In order to fully understand the variability in the predictions of the model when trained on the imbalanced data, the ARI and AMU between the confusion matrices generated when trained on uniform distribution, and each of the imbalance distributions was computed. Tables 13–15 show the obtained ARI and AMU results alongside the differences in the macro-average F1-scores for the corresponding tests in order to enable a complete analysis.

Table 13. ARI, AMU and difference in F1-scores for uniform distribution vs. equal majority–minority distribution.

Uniform Distribution vs. Equal Majority–Minority Distribution			
% Labeled Data	**ARI**	**AMU**	**Difference in F1-Scores**
80	0.14948	0.27771	0.12
60	0.28827	0.39391	0
40	0.44089	0.56003	0.06
20	0.41037	0.49083	0.04
10	0.42611	0.45006	0.10
5	0.36895	0.50927	0.18

Table 14. ARI, AMU and difference in F1-scores for uniform distribution vs. many majority distribution.

Uniform Distribution vs. Many Majority Distribution			
% Labeled Data	**ARI**	**AMU**	**Difference in F1-Scores**
80	0.30672	0.44755	0.05
60	0.20373	0.33497	0.06
40	0.26017	0.37980	0.12
20	0.36146	0.46094	0.14
10	0.24457	0.31041	0.13
5	0.36115	0.41877	0.11

Table 15. ARI, AMU and difference in F1-scores for uniform distribution vs. few majority distribution.

Uniform Distribution vs. Few Majority Distribution			
% Labeled Data	**ARI**	**AMU**	**Difference in F1-Scores**
80	0.43037	0.49422	0.02
60	0.18946	0.37004	0.08
40	0.41235	0.50588	0.17
20	0.33234	0.39484	0.07
10	0.34888	0.43448	0.04
5	0.31430	0.35841	0.11

As can be seen in this case, the equal majority–minority distribution is the most different from the uniform distribution, and this is evidenced by the low ARI score for higher amounts of labeled data. However, as the proportion of labeled data reduces, the correlation increases despite the difference in F1-scores increasing. This is indicative that, as the labeled data decreases, the model fails to adequately generalize for both distributions and, therefore, makes the same kinds of mistakes in the confusion matrices.

An interesting observation in the class containing many majority classes is that, while the ARI scores are generally constant, meaning that the correlation between the confusion matrices remains approximately the same, we can also see that the difference in F1-score remains more or less constant despite the decreasing level of labeled data beyond around 40% of the labeled samples. This suggests that, for the case of this distribution, the lowering of the proportion of labeled data has a similar effect on the model regardless of the distribution.

In the case of the few majority class distribution, it can be seen that the ARI scores indicate a significant level of correlation, with the confusion matrices generated through the uniform distribution. This could possibly be due to the few majority class distribution being the closest to the uniform distribution and, therefore, providing similar results.

Having looked at the obtained results and the correlation between the confusion matrices, this exploratory study might suggest that the FixMatch algorithm is, indeed, somewhat resistant to imbalanced data, as indicated by the ARI results. In recent times, works such as CReST [74], DARP [75], BiS [76], DASO [77] and ABC [78] have been

published, each aiming to improve the efficacy of the FixMatch algorithm in situations where the training dataset is imbalanced. While these algorithms have reported major improvements on benchmarking datasets such as CIFAR and SVHN, running each of them on this chest X-ray dataset might allow a deeper insight into how effective SSL can truly be in situations where new diseases have lower amounts of available data.

6. Conclusions and Future Work

The COVID-19 pandemic is not over yet, and people are facing difficulties due to the shortage of testing kits; hence, an alternative testing method is required CXRs can be used to detect COVID-19, as it has been seen that it could find COVID-19 in cases where the patients had symptoms, yet the PCR test returned a negative result. To reduce the burden on radiologists, deep learning can be applied for the automated detection of the disease. One limitation of the present datasets is that the data are heavily imbalanced, since the pandemic is fairly recent, and a lot of information has yet to be documented. To solve this issue, semi-supervised learning can be utilized, as these algorithms are able to generalize data with lower amounts of representative samples.

This paper therefore explored the efficacy of the state-of-the-art FixMatch semi-supervised algorithm for this problem and benchmarked the obtained results for different proportions of labeled data against supervised CNN models by utilizing transfer learning.

The results demonstrate that, even with a small proportion of labeled data, the FixMatch model is able to perform adequately well—indeed, almost along the best supervised techniques. Furthermore, an exploratory analysis was conducted toward investigating the effect of an imbalanced training dataset on the FixMatch model. The obtained results suggest that, while there is an expected drop in performance as the level of imbalance is increased, the drop is somewhat consistent and not exponential, as may be expected, thereby suggesting that the FixMatch algorithm could be somewhat robust to high levels of data imbalance.

Future investigations could explore the efficacy of other prevalent semi-supervised learning algorithms in the domain of COVID-19 detection from CXRs. Furthermore, a deeper analysis on the effects of data imbalance on the training could be carried out in order to accurately understand the behavior of the models when faced with such data. Such an experiment could lead to attempts to improve the robustness of such models for cases of high imbalance. In addition to the discussed works aiming to mitigate the effects of data imbalance, a possible class of techniques to be explored in this domain is those involving cost-based approaches being applied as part of the training process, such that the minority classes are given more importance. The enhancement of the performance of such techniques for imbalanced data will serve to improve the field in general, as such semi-supervised models will be able to generalize newer diseases for which inadequate data are present, therefore aiding medical professionals in their battle against any such outbreak.

Author Contributions: Conceptualization, I.Z.; methodology, A.R.S.; software, A.R.S.; validation, A.R.S.; formal analysis, A.R.S. and I.Z.; investigation, D.S.; resources, D.S.; data curation, A.R.S.; writing—original draft preparation, A.R.S. and D.S.; writing—review and editing, I.Z.; visualization, A.R.S.; supervision, I.Z. All authors have read and agreed to the published version of the manuscript.

Funding: The work in this paper was supported, in part, by the Open Access Program from the American University of Sharjah [grant number: OAPCEN-1410-E00064].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This paper represents the opinions of the authors and does not mean to represent the position or opinions of the American University of Sharjah.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. World Health Organization. WHO Director-General's Opening Remarks at the Media Briefing on COVID-19—11 March 2020. Available online: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 (accessed on 20 February 2022).
2. Coronavirus Worldometers. COVID Live—Coronavirus Statistics—Worldometer. Available online: https://www.worldometers.info/coronavirus/ (accessed on 21 February 2022).
3. Mayo Clinic. Coronavirus Disease 2019 (COVID-19)—Symptoms and Causes. Available online: https://www.mayoclinic.org/diseases-conditions/coronavirus/symptoms-causes/syc-20479963 (accessed on 20 February 2022).
4. Braunstein, G.D.; Schwartz, L.; Hymel, P.; Fielding, J. False Positive Results With SARS-CoV-2 RT-PCR Tests and How to Evaluate a RT-PCR-Positive Test for the Possibility of a False Positive Result. *J. Occup. Environ. Med.* **2021**, *63*, e159. [CrossRef] [PubMed]
5. Kanji, J.N.; Zelyas, N.; MacDonald, C.; Pabbaraju, K.; Khan, M.N.; Prasad, A.; Hu, J.; Diggle, M.; Berenger, B.M.; Tipples, G. False Negative Rate of COVID-19 PCR Testing: A Discordant Testing Analysis. *Virol. J.* **2021**, *18*, 13. [CrossRef] [PubMed]
6. Çallı, E.; Sogancioglu, E.; van Ginneken, B.; van Leeuwen, K.G.; Murphy, K. Deep Learning for Chest X-ray Analysis: A Survey. *Med. Image Anal.* **2021**, *72*, 102125. [CrossRef] [PubMed]
7. Shah, F.M.; Joy, S.K.S.; Ahmed, F.; Hossain, T.; Humaira, M.; Ami, A.S.; Paul, S.; Jim, A.R.K.; Ahmed, S. A Comprehensive Survey of COVID-19 Detection Using Medical Images. *SN Comput. Sci.* **2021**, *2*, 434. [CrossRef]
8. Vantaggiato, E.; Paladini, E.; Bougourzi, F.; Distante, C.; Hadid, A.; Taleb-Ahmed, A. COVID-19 Recognition Using Ensemble-CNNs in Two New Chest X-ray Databases. *Sensors* **2021**, *21*, 1742. [CrossRef]
9. Yang, X.; Song, Z.; King, I.; Xu, Z. A Survey on Deep Semi-Supervised Learning. *arXiv* **2021**, arXiv:2103.00550.
10. Ouali, Y.; Hudelot, C.; Tami, M. An Overview of Deep Semi-Supervised Learning. *arXiv* **2020**, arXiv:2006.05278.
11. Sohn, K.; Berthelot, D.; Li, C.L.; Zhang, Z.; Carlini, N.; Cubuk, E.D.; Kurakin, A.; Zhang, H.; Raffel, C. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *arXiv* **2020**, arXiv:2001.07685.
12. Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Ball, R.L.; Langlotz, C.; et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-rays with Deep Learning. *arXiv* **2017**, arXiv:1711.05225.
13. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269.
14. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. ChestX-ray8: Hospital-Scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3462–3471.
15. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
16. Baltruschat, I.M.; Nickisch, H.; Grass, M.; Knopp, T.; Saalbach, A. Comparison of Deep Learning Approaches for Multi-Label Chest X-ray Classification. *Sci. Rep.* **2019**, *9*, 6381. [CrossRef]
17. Irfan, A.; Adivishnu, A.L.; Sze-To, A.; Dehkharghanian, T.; Rahnamayan, S.; Tizhoosh, H.R. Classifying Pneumonia among Chest X-rays Using Transfer Learning. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 20–24 July 2020; pp. 2186–2189.
18. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
19. Hwang, S.; Kim, H.E.; Jeong, J.; Kim, H.J. A Novel Approach for Tuberculosis Screening Based on Deep Convolutional Neural Networks. In *Medical Imaging 2016: Computer-Aided Diagnosis*; Tourassi, G.D., Armato, S.G., Eds.; SPIE: San Diego, CA, USA, 2016; pp. 750–757.
20. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2012; Volume 25.
21. Liu, C.; Cao, Y.; Alcantara, M.; Liu, B.; Brunette, M.; Peinado, J.; Curioso, W. TX-CNN: Detecting Tuberculosis in Chest X-ray Images Using Convolutional Neural Network. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 2314–2318.
22. Rahman, T.; Khandakar, A.; Kadir, M.A.; Islam, K.R.; Islam, K.F.; Mazhar, R.; Hamid, T.; Islam, M.T.; Kashem, S.; Mahbub, Z.B.; et al. Reliable Tuberculosis Detection Using Chest X-ray with Deep Learning, Segmentation and Visualization. *IEEE Access* **2020**, *8*, 191586–191601. [CrossRef]
23. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, Germany, 5–9 October 2015; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
24. Amin, I.; Hassan, S.; Jaafar, J. Semi-Supervised Learning for Limited Medical Data Using Generative Adversarial Network and Transfer Learning. In Proceedings of the 2020 International Conference on Computational Intelligence (ICCI), Bandar Seri Iskandar, Malaysia, 8–9 October 2020; pp. 5–10.
25. Sajun, A.R.; Zualkernan, I. Survey on Implementations of Generative Adversarial Networks for Semi-Supervised Learning. *Appl. Sci.* **2022**, *12*, 1718. [CrossRef]

26. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015; Bengio, Y., LeCun, Y., Eds.; Conference Track Proceedings.
27. Zhang, W.; Wang, H.; Lai, Z.; Hou, C. Constrained Contrastive Representation: Classification on Chest X-rays with Limited Data. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; pp. 1–6.
28. Oh, Y.; Park, S.; Ye, J.C. Deep Learning COVID-19 Features on CXR Using Limited Training Data Sets. *IEEE Trans. Med. Imaging* **2020**, *39*, 2688–2700. [CrossRef] [PubMed]
29. Mangal, A.; Kalia, S.; Rajgopal, H.; Rangarajan, K.; Namboodiri, V.; Banerjee, S.; Arora, C. COVIDAID: COVID-19 Detection Using Chest X-ray. *arXiv* **2020**, arXiv:2004.09803.
30. Apostolopoulos, I.D.; Mpesiana, T.A. COVID-19: Automatic Detection from X-ray Images Utilizing Transfer Learning with Convolutional Neural Networks. *Phys. Eng. Sci. Med.* **2020**, *43*, 635–640. [CrossRef] [PubMed]
31. Ozturk, T.; Talo, M.; Yildirim, E.A.; Baloglu, U.B.; Yildirim, O.; Rajendra Acharya, U. Automated Detection of COVID-19 Cases Using Deep Neural Networks with X-ray Images. *Comput. Biol. Med.* **2020**, *121*, 103792. [CrossRef]
32. Narin, A.; Kaya, C.; Pamuk, Z. Automatic Detection of Coronavirus Disease (COVID-19) Using X-ray Images and Deep Convolutional Neural Networks. *Pattern Anal. Appl.* **2021**, *24*, 1207–1220. [CrossRef]
33. Khasawneh, N.; Fraiwan, M.; Fraiwan, L.; Khassawneh, B.; Ibnian, A. Detection of COVID-19 from Chest X-ray Images Using Deep Convolutional Neural Networks. *Sensors* **2021**, *21*, 5940. [CrossRef]
34. Luz, E.; Silva, P.L.; Silva, R.; Silva, L.; Moreira, G.; Menotti, D. Towards an Effective and Efficient Deep Learning Model for COVID-19 Patterns Detection in X-ray Images. *Res. Biomed. Eng.* **2021**, *38*, 149–162. [CrossRef]
35. AbdElhamid, A.A.; AbdElhalim, E.; Mohamed, M.A.; Khalifa, F. Multi-Classification of Chest X-rays for COVID-19 Diagnosis Using Deep Learning Algorithms. *Appl. Sci.* **2022**, *12*, 2080. [CrossRef]
36. Al-Shargabi, A.A.; Alshobaili, J.F.; Alabdulatif, A.; Alrobah, N. COVID-CGAN: Efficient Deep Learning Approach for COVID-19 Detection Based on CXR Images Using Conditional GANs. *Appl. Sci.* **2021**, *11*, 7174. [CrossRef]
37. Wang, L.; Lin, Z.Q.; Wong, A. COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-ray Images. *Sci. Rep.* **2020**, *10*, 19549. [CrossRef] [PubMed]
38. Arias-Londono, J.D.; Gomez-Garcia, J.A.; Moro-Velazquez, L.; Godino-Llorente, J.I. Artificial Intelligence Applied to Chest X-ray Images for the Automatic Detection of COVID-19. A Thoughtful Evaluation Approach. *IEEE Access* **2020**, *8*, 226811–226827. [CrossRef] [PubMed]
39. Hertel, R.; Benlamri, R. COV-SNET: A Deep Learning Model for X-ray-Based COVID-19 Classification. *Inform. Med. Unlocked* **2021**, *24*, 100620. [CrossRef] [PubMed]
40. Sahlol, A.T.; Yousri, D.; Ewees, A.A.; Al-qaness, M.A.A.; Damasevicius, R.; Elaziz, M.A. COVID-19 Image Classification Using Deep Features and Fractional-Order Marine Predators Algorithm. *Sci. Rep.* **2020**, *10*, 15364. [CrossRef]
41. Kedia, P.; Katarya, R. CoVNet-19: A Deep Learning Model for the Detection and Analysis of COVID-19 Patients. *Appl. Soft Comput.* **2021**, *104*, 107184. [CrossRef]
42. Mahanty, C.; Kumar, R.; Asteris, P.G.; Gandomi, A.H. COVID-19 Patient Detection Based on Fusion of Transfer Learning and Fuzzy Ensemble Models Using CXR Images. *Appl. Sci.* **2021**, *11*, 11423. [CrossRef]
43. Win, K.Y.; Maneerat, N.; Sreng, S.; Hamamoto, K. Ensemble Deep Learning for the Detection of COVID-19 in Unbalanced Chest X-ray Dataset. *Appl. Sci.* **2021**, *11*, 10528. [CrossRef]
44. Haque, S.; Hoque, M.A.; Khan, M.A.I.; Ahmed, S. COVID-19 Detection Using Feature Extraction and Semi-Supervised Learning from Chest X-ray Images. In Proceedings of the 2021 IEEE Region 10 Symposium (TENSYMP), Jeju, Korea, 23–25 August 2021; pp. 1–5.
45. Berthelot, D.; Carlini, N.; Goodfellow, I.; Oliver, A.; Papernot, N.; Raffel, C. MixMatch: A Holistic Approach to Semi-Supervised Learning. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 13–14 December 2019; Curran Associates Inc.: Red Hook, NY, USA, 2019.
46. Calderon-Ramirez, S.; Giri, R.; Yang, S.; Moemeni, A.; Umana, M.; Elizondo, D.; Torrents-Barrena, J.; Molina-Cabello, M.A. Dealing with Scarce Labelled Data: Semi-Supervised Deep Learning with Mix Match for COVID-19 Detection Using Chest X-ray Images. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 5294–5301.
47. Calderon-Ramirez, S.; Yang, S.; Moemeni, A.; Elizondo, D.; Colreavy-Donnelly, S.; Chavarría-Estrada, L.F.; Molina-Cabello, M.A. Correcting Data Imbalance for Semi-Supervised COVID-19 Detection Using X-ray Chest Images. *Appl. Soft Comput.* **2021**, *111*, 107692. [CrossRef]
48. Abbas, A.; Abdelsamea, M.M.; Gaber, M.M. 4S-DT: Self-Supervised Super Sample Decomposition for Transfer Learning with Application to COVID-19 Detection. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 2798–2808. [CrossRef]
49. Gazda, M.; Plavka, J.; Gazda, J.; Drotar, P. Self-Supervised Deep Convolutional Neural Network for Chest X-ray Classification. *IEEE Access* **2021**, *9*, 151972–151982. [CrossRef]
50. Oliver, A.; Odena, A.; Raffel, C.; Cubuk, E.D.; Goodfellow, I.J. Realistic Evaluation of Deep Semi-Supervised Learning Algorithms. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; Curran Associates Inc.: Red Hook, NY, USA, 2018; pp. 3239–3250.
51. Van Engelen, J.E.; Hoos, H.H. A Survey on Semi-Supervised Learning. *Mach. Learn.* **2020**, *109*, 373–440. [CrossRef]

52. Lee, D.H. Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. In Proceedings of the Workshop on Challenges in Representation Learning, Atlanta, GA, USA, 16–21 June 2013; Volume 3.
53. Tarvainen, A.; Valpola, H. Mean Teachers Are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 1195–1204.
54. Xie, Q.; Dai, Z.; Hovy, E.; Luong, M.-T.; Le, Q.V. Unsupervised Data Augmentation for Consistency Training. *arXiv* **2020**, arXiv:1904.12848.
55. Berthelot, D.; Carlini, N.; Cubuk, E.D.; Kurakin, A.; Sohn, K.; Zhang, H.; Raffel, C. ReMixMatch: Semi-Supervised Learning with Distribution Alignment and Augmentation Anchoring. *arXiv* **2020**, arXiv:1911.09785.
56. Krizhevsky, A. *Learning Multiple Layers of Features from Tiny Images*; Technical Report; University of Toronto: Toronto, ON, Canada, 2009.
57. Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; Ng, A.Y. Reading Digits in Natural Images with Unsupervised Feature Learning. In Proceedings of the NIPS Workshop on Deep Learning and Unsupervised Feature Learning, Granada, Spain, 12–17 December 2011.
58. Cohen, J.P.; Morrison, P.; Dao, L.; Roth, K.; Duong, T.Q.; Ghassemi, M. COVID-19 Image Data Collection: Prospective Predictions Are the Future. *arXiv* **2020**, arXiv:2006.11988.
59. Kermany, D.S.; Goldbaum, M.; Cai, W.; Valentim, C.C.S.; Liang, H.; Baxter, S.L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell* **2018**, *172*, 1122–1131.e9. [CrossRef]
60. Shih, G.; Wu, C.C.; Halabi, S.S.; Kohli, M.D.; Prevedello, L.M.; Cook, T.S.; Sharma, A.; Amorosa, J.K.; Arteaga, V.; Galperin-Aizenberg, M.; et al. Augmenting the National Institutes of Health Chest Radiograph Dataset with Expert Annotations of Possible Pneumonia. *Radiol. Artif. Intell.* **2019**, *1*, e180041. [CrossRef]
61. Jaeger, S.; Candemir, S.; Antani, S.; Wáng, Y.X.; Lu, P.X.; Thoma, G. Two Public Chest X-ray Datasets for Computer-Aided Screening of Pulmonary Diseases. *Quant. Imaging Med. Surg.* **2014**, *4*, 475–477. [CrossRef]
62. PyTorch. Available online: https://www.pytorch.org (accessed on 18 August 2021).
63. Kekmodel/FixMatch-Pytorch at F54946074fba383e28320d8f50b627eabd0c7e3c. Available online: https://github.com/kekmodel/FixMatch-pytorch (accessed on 23 August 2021).
64. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
65. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807.
66. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
67. Tan, M.; Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; Chaudhuri, K., Salakhutdinov, R., Eds.; PMLR. Volume 97, pp. 6105–6114.
68. Rinne, H. *The Weibull Distribution: A Handbook*; CRC Press: Boca Raton, FL, USA, 2008.
69. Weibull Distribution Applet/Calculator. Available online: https://homepage.divms.uiowa.edu/~{}mbognar/applets/weibull.html (accessed on 16 January 2022).
70. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
71. Lin, J. Divergence Measures Based on the Shannon Entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151. [CrossRef]
72. Maćkiewicz, A.; Ratajczak, W. Principal Components Analysis (PCA). *Comput. Geosci.* **1993**, *19*, 303–342. [CrossRef]
73. Van der Maaten, L.; Hinton, G. Visualizing Data Using T-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
74. Wei, C.; Sohn, K.; Mellina, C.; Yuille, A.; Yang, F. CReST: A Class-Rebalancing Self-Training Framework for Imbalanced Semi-Supervised Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 10857–10866.
75. Kim, J.; Hur, Y.; Park, S.; Yang, E.; Hwang, S.J.; Shin, J. Distribution Aligning Refinery of Pseudo-Label for Imbalanced Semi-Supervised Learning. In *Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–12 December 2020*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 14567–14579.
76. He, J.; Kortylewski, A.; Yang, S.; Liu, S.; Yang, C.; Wang, C.; Yuille, A. Rethinking Re-Sampling in Imbalanced Semi-Supervised Learning. *arXiv* **2021**, arXiv:2106.00209.
77. Oh, Y.; Kim, D.J.; Kweon, I.S. Distribution-Aware Semantics-Oriented Pseudo-Label for Imbalanced Semi-Supervised Learning. *arXiv* **2021**, arXiv:2106.05682.
78. Lee, H.; Shin, S.; Kim, H. ABC: Auxiliary Balanced Classifier for Class-Imbalanced Semi-Supervised Learning. *arXiv* **2021**, arXiv:2110.10368.