Semi/Self-Supervised Learning on a Pediatric Pneumonia Dataset

Wing Poon, Sundeep Bhimireddy, Sinem Erisken

FourthBrain Capstone Project Technical Report

1. Background and Significance

One of the leading causes of death in children is pediatric pneumonia. Pediatric pneumonia, according to the World Health Organization (WHO), killed nearly 750,000 children worldwide in 2019, accounting for 14% of all deaths of children under the age of 5 globally. As a global crisis, pediatric pneumonia most severely threatens developing countries. Countries such as Bangladesh, Zambia, India, Kenya, and Uganda have developed strict plans to control pneumonia on both the national and state level, as reported by the WHO. As many children can recover without the need for treatment, urgent cases can be missed or left untreated. Fortunately, treatment of even critical cases can be as simple as administering antibiotics and diagnosis often only relies on the common chest x-ray.

As physician time is valuable and expensive, there exists a huge opportunity for artificial intelligence (AI), and particularly deep learning (DL), to bring urgent cases to the attention of the attending physician, thereby potentially saving lives. Unfortunately, the current state-of-the-art (SOTA) methods detecting pediatric pneumonia use fully supervised learning that needs large and fully annotated datasets. Hence, only hospitals which can afford to both collect and annotate large datasets can actually use these methods to aid their physicians.

Our primary goal in this project is to reduce the labeling cost, and hence democratize AI applications for the front lines of pediatric pneumonia. Instead of a fully-supervised approach, here we employ both self and semi-supervised methods in order to dramatically reduce the number of required labels while minimally sacrificing accuracy. Importantly, while our semi-supervised approach did not outperform our supervised baseline, it nonetheless did outperform many SOTA reports.

2. Related Work

The problem of using AI and DL to diagnose pediatric pneumonia is certainly not new and has been extensively reviewed elsewhere (Kahn et al. 2021; see also Kermany et al. 2018; Labhane et al. 2020). Current SOTA approaches have relied on fully supervised learning.

Table 1 highlights results from select previous studies who have looked at the same dataset (see section 3 for description of the dataset). Deep transfer learning, most commonly transferred from ImageNet, is a very popular SOTA approach (Kernamy et al. 2018; Labhane et al. 2020; Hashmi et al. 2020). Even so, we decided against transfer learning. As our method (see section 4) relies on robust representations of lung x-rays created during the first stage of training, we instead opted to train our network from scratch.

Publication	Metric	Results
Kermany et al. 2018	Accuracy	92%
Stephen et al. 2019	Validation Accuracy	93.73%
Labhane et al. 2020	Accuracy	97%, 98%
Hashmi et al. 2020	Accuracy	98.43%

Table 1: SOTA results from previous publications.
Table modified from Labhane et al. 2020. Results from Labhane et al. 2020 have been condensed to include results from varying architectures. Results from Hashmi et al. 2020 only include their best and most relevant report of using a weighted classifier with optimized weights across different CNN architectures.

Labhane and colleagues (Labhane et al. 2020), as well as Hashmi and colleagues (Hashmi et al. 2020), explored different model architectures, including InceptionV3 and Xception. While Labhane and colleagues found their best results using InceptionV3, Hashmi and colleagues found optimal performance using a weighted classifier which encompassed multiple architectures. Due to time constraints, model architecture was a hyperparameter we only minimally tuned within different Resnets, so we can not speak to the advantages of our chosen architecture in comparison to CNNs like Xception and Inception except to say that our supervised baseline outperformed these SOTA reports (see section 5, compare Table 1 and Table 2). Additionally, we believe a weighted classifier approach is counter productive to our project goal of reducing the cost of an AI application, as a weighted architecture requiring multiple parallel CNNs greatly increases already expensive computing costs.

3. Explanation of dataset

The dataset we used is previously published (Kermany et al. 2018), publicly available, and consists of 5855 monochromatic images of pediatric lung x-rays. Upon downloading, the dataset is already split into a train (5232 images) and a test (623 images) set. The dataset is unbalanced: the pathological condition compared to the normal condition is greatly overrepresented in both the train (74%:26%) and test (62%:38%) splits. Also consider that although pneumonia is overrepresented in both splits, the data is differently distributed in the original test and train splits. The dataset is also highly variable in image properties such as resolution and aspect ratio. Basic preprocessing of the images has been performed to reduce the variance in aspect ratio. In an effort to homogenize train, validation and test distributions, we performed our own splits of the data after reshuffling. Reshuffling also seems to be a main methodological difference between previous SOTA publications which performed >95% and those which did not in Table 1.

4. Methods

SimCLR(v2) is a powerful self/semi-supervised learning technique (Chen et al. 2020) with a publicly available code repository on github that is compatible with TensorFlow 2. We modified existing code, where necessary, to be able to conduct model training on the selected dataset and model architectures. We trained the model on Google Cloud Platform using 8 TPUs.

4.1 Data preprocessing

4.2 ML Workflow

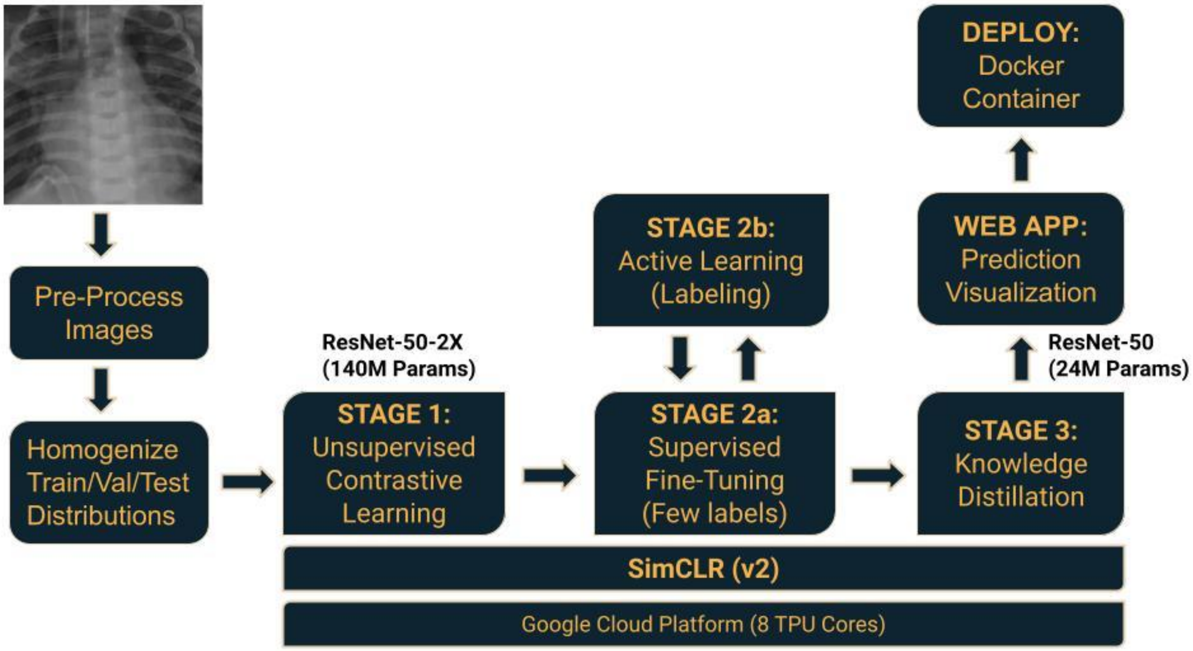

Pre-Process Images
Homogenize Train/Val/Test Distributions
ResNet-50-2X (140M Params)
STAGE 1: Unsupervised Contrastive Learning
STAGE 2b: Active Learning (Labeling)
STAGE 2a: Supervised Fine-Tuning (Few labels)
STAGE 3: Knowledge Distillation
ResNet-50 (24M Params)
WEB APP: Prediction Visualization
DEPLOY: Docker Container
SimCLR (v2)
Google Cloud Platform (8 TPU Cores)

Figure 1: ML Workflow

SimCLR(v2) consists of three stages: 1) Stage 1- Unsupervised learning 2) Stage 2 - supervised fine-tuning 3) Stage 3 - distillation with unlabeled images.

4.2.1. Stage 1: Unsupervised representation learning

The first stage involves learning general visual representations with unlabeled images. The representations are learned by maximizing agreement between the differently augmented views of the same image using contrastive loss objective. For example, given an image sample, the image is augmented twice using random crop, color distortion and Gaussian blur generating two images of the same sample. And then the model is trained on these augmented images by minimizing the NT-Xent loss function. The general representation learning benefits from having bigger and wider models, therefore, we have chosen a wide ResNet-50 (2X width with Selective

Kernels) which consists of 140M trainable parameters for the Stage 1 pre-training and Stage 2 fine-tuning.

The standard Inception-style random cropping (crop of random size 0.08-1.0 in area of the original size) yielded poorer results for this particular dataset because the ROI is much smaller compared to the ImageNet dataset. Hypertuning on this parameter has resulted in 0.5-1.0 crop area as optimal range for the x-ray dataset. Also, we used LARS optimizer with a momentum of 0.9, cosine decay schedule, weight decay of $1e^{-4}$, learning rate of 0.2, and temperature of 0.1. We trained the model with a batch size of 128 for a total of 500 epochs reaching a contrastive accuracy of 99.6%.

4.2.2. Stage 2: Semi-supervised fine tuning and active learning

For fine-tuning, we attached the linear classifier to the first layer of the projection head and fine-tuned all the layers for 1%, 2%, 5% labeled examples without any cropping. The labels were chosen using two approaches: 1) random 2) active learning (augmentation based). We found that both approaches performed similarly for the 1% and 2% labeled scenarios. For the 5% labeled scenario, we found that the active-learning approach performed significantly better than random selection. Stage 2 fine-tuned model results are shown in Table 2 below. We trained the models for 60 epochs with a batch size of 32, and with a learning rate of 0.001.

4.2.3. Stage 3: Knowledge distillation

For distillation, we only used unlabeled examples and distilled the knowledge from the Stage 2 fine-tuned wide ResNet-50 model (teacher) to a standard ResNet-50 model (student). The student model which has 24M trainable parameters is 5X-6X smaller than the teacher model. We set the temperature to 0.1, and used the same learning rate schedule, weight decay, batch size as Stage 1 pre training. Table 2 below shows the Stage 3 model performance by the label %. Like in Stage 2, active learning based label selection performed better over random selection for the 5% label scenario.

5. Results

Table 2 summarizes our results on the test dataset. SimCLR models from both Stage 2 and 3 consistently outperformed Fully Supervised Learning (FSL) model performance in all three scenarios. In Stage 3, not only were we able to decrease the size of the model (making it more suitable for deployment), but also were able to improve the model performance consistently. Compared to the FSL model performance (98.9% top-1 accuracy) for the 100% label scenario, the Stage 3 distilled models were able to achieve a top-1 accuracy of 97.6% and 98.1% using only 2% and 5% labels respectively.

Labels (#)	Labels (%)	Fully Supervised	FixMatch	SimCLR (v2)	
				Stage 2 (Fine-tuning)	Stage 3 (Distillation)
52	1%	85.2	92.1	94.5	96.3
104	2%	87.2	95.0	96.8	**97.6**
260	5%	86.0	98.2	97.1*	**98.1***
4708	100%	**98.9**	-	-	-

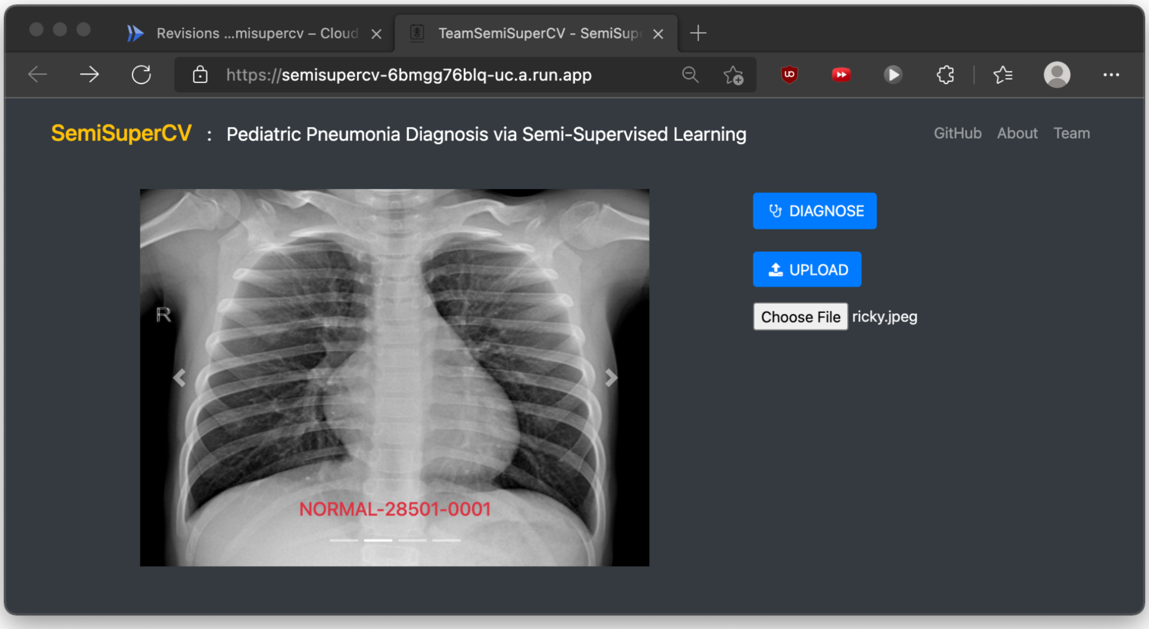
Table 2: Classification Accuracies
The most relevant results: our fully supervised baseline and our distilled model using 2% and 5% of labels, are highlighted in bold. * indicates that we used Active Learning to pick our labels

In addition to accuracy, we also looked at the precision and recall for the 2% and 5% labeled Stage 3 models. The 2% labeled Stage 3 model had a precision of 97.8% and a recall of 98.9%. The 5% labeled Stage 3 model had a precision of 98.5% and a recall of 98.9%.

6. Deployment

Trained and distilled down to a small and fast CNN with just 24M parameters, the model was suitable for production deployment. We created a web application to illustrate how users can interact with the model in real-time.

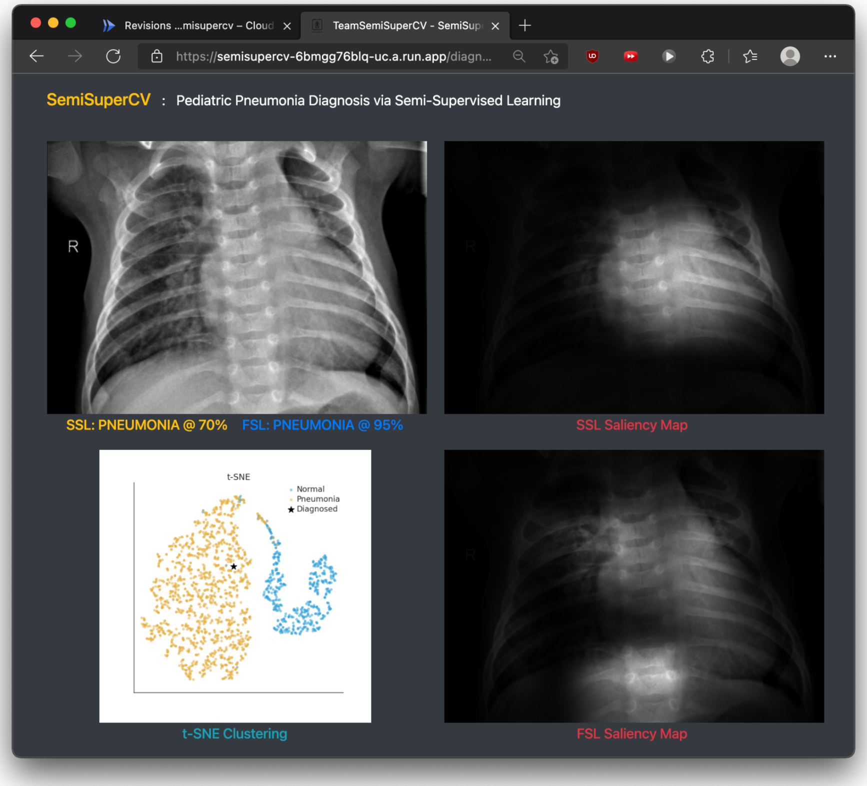
6.1 Web Application



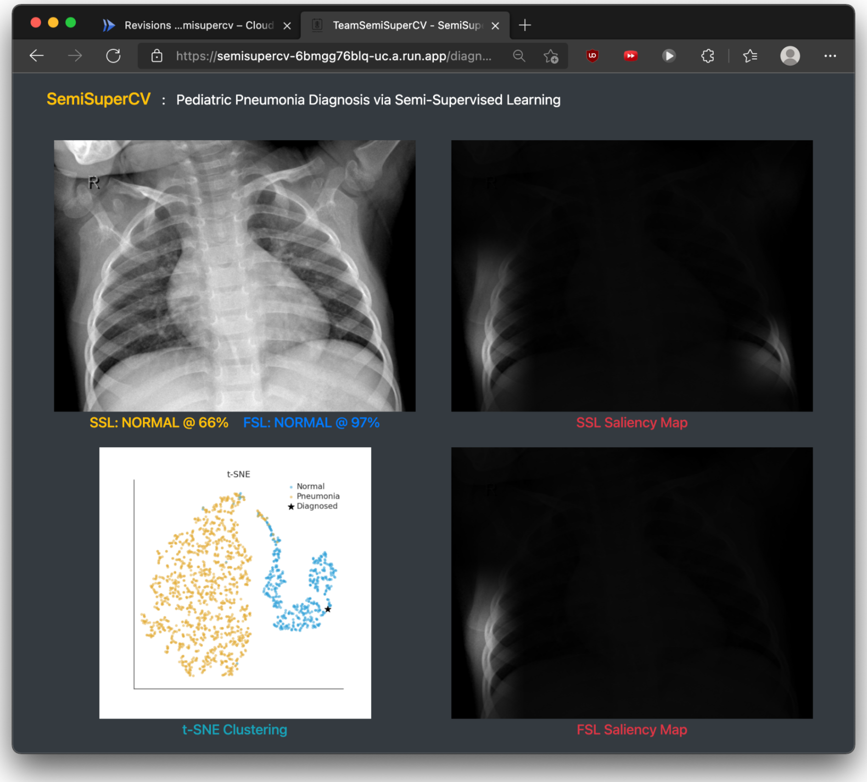
https://semisupercv-6bmgg76blq-uc.a.run.app

SemiSuperCV : Pediatric Pneumonia Diagnosis via Semi-Supervised Learning

Choose File ricky.jpeg



SemiSuperCV : Pediatric Pneumonia Diagnosis via Semi-Supervised Learning



SemiSuperCV : Pediatric Pneumonia Diagnosis via Semi-Supervised Learning

Figure 2: Front page of deployed web app
Top: Front Page. *Middle:* Pneumonia diagnosis with saliency maps and t-SNE plot. *Bottom:* Normal diagnosis with saliency maps and t-SNE plot.

The application was written using FastAPI for implementing the REST API backend, HTML, Bootstrap and Javascript for the front-end, and Tensorflow-CPU for inference. It was containerized using Docker and deployed to the Google Cloud Platform. It can be invoked using Cloud Run for serverless deployment scalability.

6.2 Saliency maps

Deep Learning models are often criticised for being a 'Black Box', which can be a problem for medical applications. With the help of modern visualization techniques, we sought to alleviate this concern. The app shows the saliency maps of two competing models we trained. Each map is obtained by probing the activation of the last CNN layer of the model to highlight areas of the submitted image that produced the greatest change (i.e. gradient) in activation magnitudes.

The top map is from our semi-supervised (SS) model, while the bottom is from our fully supervised (FS) model. We presented both to allow comparison of whether our SS model 'pays attention' to the same regions of the image as that of a traditionally-trained FS model. Illuminated (i.e. non-masked) areas indicate the most salient regions for each model. We note that there is generally good correspondence between the models. For example, in cases with pneumonia, both models were 'focussing' squarely at the lungs.

6.3 t-distributed stochastic neighbor embedding (t-SNE)

On the bottom left of the app, we show a t-SNE visualization of the SS model's embedding space. A t-SNE plot was chosen because it is a non-parametric, non-linear method of projecting high-dimensional space data into low dimensions, in this case a 2D graph, while preserving clusters of neighboring points in the original embedding. For context, we first t-SNE transformed 2000 Training examples from our dataset, using the latent representation called from the last dense layer of the model. The color of each point denotes the true label of that example -- blue for Normal, and orange for Pneumonia. We then placed a Star to indicate the t-SNE mapping of the sample that is being diagnosed. We observed that the stars from the Test set generally fall in the cluster of points that are associated with each classification, with predictions of low confidence falling in the 'causeway' linking the two clusters.

7. Ethical Considerations

7.1 Informed Consent to Use

The data collection was conducted in a manner compliant with the United States Health Insurance Portability and Accountability Act (HIPAA) and was adherent to the tenets of the Declaration of Helsinki[2].

7.2 Safety

The software function is intended *for the purpose of enabling such healthcare professional to independently review the basis for such recommendations that such software presents so that it is not the intent that such healthcare professionals rely primarily on any of such recommendations to make a clinical diagnosis or treatment decision regarding an individual patient*[1].

7.3 Transparency

The source of data used in our study is from: Large Dataset of Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images - Mendeley Data

Chest X-ray images (anterior-posterior) were selected from retrospective cohorts of pediatric patients of one to five years old from Guangzhou Women and Children's Medical Center, Guangzhou. Institutional Review Board (IRB)/Ethics Committee approvals were obtained. The work was conducted in a manner compliant with the United States Health Insurance Portability and Accountability Act (HIPAA) and was adherent to the tenets of the Declaration of Helsinki[2].

[1] [Clinical Decision Support Software, Draft Guidance for Industry and FDA Staff]

[2] [Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning]

7.4 Algorithmic Bias

X-Ray imaging is a common, inexpensive diagnosis tool and it is unlikely that the distribution of patients receiving these scans will exhibit socio-economic demographic skew. All imaging was performed as part of patients' routine clinical care.

7.5 Data Privacy

All data samples are fully anonymous with no personally-identifying information (e.g. facial bone structure).

7.6 Unintended Consequences

It is possible that advanced techniques for semi-supervised learning will allow for more advanced citizen surveillance. For example, increased efficacy of few-shot classification might allow for more accurate person identification from just a few images. The intended use of the model is to automatically surface suspected urgent cases for expedited diagnosis by a radiologist or clinician and not to replace expert medical opinion.

8. Summary

We were successful in developing, validating and deploying a Semi-Supervised Learning model that achieved a classification accuracy of almost 98% with just 2% of the dataset labeled. We did this using the SimCLR Contrastive Loss Representational Learning framework, in conjunction with Knowledge Distillation and Active Learning. Despite such sparse annotations, we show performance greatly eclipsing that of standard Fully-Supervised Learning.

We found empirically that hyper-parameter tuning using a validation dataset is necessary for guiding on the path towards good final model performance, hence the importance of remixing the Train/Val/Test splits to ensure that the validation split is predictive of the whole.

We discovered that, unlike ImageNet, our dataset required constraining the random cropping augmentations that is used for Consistency Regularization to not be too small, as the pneumonia infiltrates comprise only a small portion of the entire image frame.

Finally, we also tried an older semi-supervised learning technique, FixMatch, but albeit using only the framework-default settings and without time to perform hyper-parameter tuning, we achieved better results with SimCLR (with <5% annotations).

8. Future Directions

1. Try more aggressive augmentations during the Stage 2 Fine-Tuning step to improve the active labeling strategy (by making 'hard' cases harder yet).
2. To reduce the need for validation examples, we want to explore the generalizability of our chosen hyperparameters to other medical datasets.
3. Having tried both entropy as well as augmentation-based Active-Learning algorithms, with neither being clear winners, we are curious to try other label selection algorithms.
4. Quoting Yann LeCun, "Self-Supervised Learning is one of the most promising ways of approximating a form of common sense in AI". The field is rapidly evolving and new frameworks are emerging at a rapid clip so we should try other emergent SOTA techniques.

References

1. Chen, Ting, et al. "Big self-supervised models are strong semi-supervised learners." *arXiv preprint arXiv:2006.10029* (2020).
2. Hashmi, Mohammad Farukh, et al. "Efficient pneumonia detection in chest xray images using deep transfer learning." *Diagnostics* 10.6 (2020): 417.
3. Kermany, Daniel S., et al. "Identifying medical diagnoses and treatable diseases by image-based deep learning." *Cell* 172.5 (2018): 1122-1131.
4. Khan, Wasif, Nazar Zaki, and Luqman Ali. "Intelligent pneumonia identification from chest x-rays: A systematic literature review." *IEEE Access* (2021).
5. Labhane, Gaurav, et al. "Detection of pediatric pneumonia from chest x-ray images using cnn and transfer learning." *2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE)*. IEEE, 2020.
6. Stephen, Okeke, et al. "An efficient deep learning approach to pneumonia classification in healthcare." *Journal of healthcare engineering* 2019 (2019).